INCORPORACIÓN DE TÉCNICAS MULTIVARIANTES EN UN SISTEMA GESTOR DE BASES DE DATOS

TESIS DE MAESTRÍA

Autoría de:
CARLOS MARIO SOTO JARAMILLO

Directora:
Ph. D. CLAUDIA JIMÉNEZ RAMÍREZ

MAESTRÍA EN INGENIERÍA - INGENIERÍA DE SISTEMAS
FACULTAD DE MINAS

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
2009

AGRADECIMIENTOS

Por el esfuerzo, apoyo y dedicación, presento mis más sinceros agradecimientos a la Profesora Asociada, Claudia Jiménez Ramírez, adscrita a la Facultad de Minas de la Universidad Nacional de Colombia, Sede Medellín, quien como directora de la Tesis de Maestría, merece mi respeto y admiración por su apoyo en la lectura cuidadosa, sugerencias y aportes presentados.

Adicionalmente, quiero ofrecer mis agradecimientos a la Facultad de Minas, quien me otorgó una Beca durante mis estudios de Maestría, sintiéndome satisfecho de haber continuado estudios de posgrado en la Universidad Nacional de Colombia, reconocida por su calidad académica e institucional.

Por último, se agradece a todas las personas que de una u otra forma han participado o colaborado con su conocimiento en el desarrollo del trabajo investigativo. Igualmente, a mi esposa Alejandra Restrepo Castañeda, y demás familiares, por su apoyo y comprensión.

RESUMEN

El objetivo principal de la presente Tesis de Maestría es la incorporación de las técnicas de regresión lineal y logística multivariante en un sistema gestor de bases de datos, con el propósito de facilitar el Descubrimiento de Conocimientos en Bases de Datos y promover el enfoque de Inteligencia del Negocio con una herramienta con la inteligencia suficiente para interpretar y presentar los resultados de manera amigable para apoyar la toma de decisiones.

Se propuso un modelo conceptual para la incorporación de las técnicas multivariantes en un sistema gestor de bases de datos, adicionalmente, se presenta un modelo para la visualización de resultados y se desarrolló un prototipo de una aplicación Web para verificar la factibilidad técnica del modelo propuesto.

Se muestra cómo el modelo propuesto para la visualización de los resultados posee la potencia expresiva para facilitar la asimilación del nuevo conocimiento generado con el análisis de regresión. Se demostró que el prototipo desarrollado facilita la selección de los datos para un análisis de regresión e interpreta por sí mismo los resultados, facilitando el Descubrimiento de Conocimiento en Bases de Datos a usuarios no expertos. Finalmente, el modelo conceptual para la incorporación de las técnicas de regresión multivariantes en un sistema gestor de bases de datos y el modelo para la visualización de los resultados, presentan las características apropiadas para brindar soporte a proyectos de Descubrimiento de Conocimiento en Bases de Datos.

ABSTRACT

The principal objective of this Thesis is to incorporate techniques of multivariate linear and logistic regressions in a database management system, with the purpose of facilitate Knowledge Discovery in Databases and to promote the Business Intelligence with an intelligent tool that allows to interpret and to present the results in friendly way to support the decision making.

A conceptual model which incorporates techniques related with multivariate regressions in database management system is proposed. Additionally, a model for the visualization of results is presented and a Web application prototype to verify the technical feasibility of the model is developed.

The visualization model of regression results showed expressive power to facilitate the assimilation of new knowledge produced with regression analysis. It was demonstrated that the developed prototype facilitates the selection of data for a regression analysis and interprets the results by itself, making easy to non expert users the Knowledge Discovery in Databases. Finally, the conceptual model for incorporating techniques of multivariate regressions in a database management system and the model for the visualization of results presented characteristics appropriated to support the projects to Knowledge Discovery in Databases.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

En la actualidad el mundo es cada vez más competitivo, por esto, la toma de decisiones debe ser más acertada y oportuna para tener una mayor capacidad de lograr un efecto determinado, o ser capaz de reaccionar rápidamente, adaptándose a nuevas circunstancias. Todo esto implica, la necesidad de un fortalecimiento de los sistemas informáticos para apoyar la toma de decisiones y la necesidad del análisis de los datos disponibles en bases de datos u otros. Lo anterior explica la tendencia creciente en la utilización de la información almacenada para apoyar decisiones administrativas; pero aún, un gran número de organizaciones desconocen los beneficios del Descubrimiento de Conocimiento en Bases de Datos. Una de las razones puede ser que no se valore el impacto real de la carencia de buena información sobre la rentabilidad del negocio. Otra razón, puede ser la ausencia de un equipo de desarrollo dentro de la empresa, o quizás en algunos casos, se encuentran organizaciones cuyos ejecutivos están poco preparados para realizar estas labores técnicas.

El manejo de la información es una actividad diaria en cualquier organización, empresa o institución. Un primer paso para facilitar el manejo de los datos generados por las operaciones diarias son las bases de datos operacionales. Las bases de datos operacionales surgieron hace algunas décadas, facilitando el almacenamiento de la información y siendo diseñadas para optimizar o facilitar el trabajo cotidiano, que le sirve a la empresa para realizar sus operaciones básicas.

Actualmente, el almacenamiento de la información es sencillo y económico; los gestores de bases de datos son las principales herramientas para almacenar grandes cantidades de información. Estas herramientas ofrecen una interfaz o lenguaje interactivo para plantear solicitudes de información. La mayoría de los gestores de bases de datos convencionales soportan el lenguaje estructurado de búsqueda SQL (por sus siglas en inglés: *Structured Query Language*). SQL ha demostrado ser bastante versátil en la recuperación de información en bases de datos, convirtiéndose en el lenguaje de consulta más popular. Los sistemas de consulta-respuesta basados en SQL pueden responder a una gran cantidad de preguntas que representan gran parte del conocimiento almacenado en una base de datos; sin embargo, existe conocimiento oculto, representado en parte por relaciones que no son identificables a simple vista.

Extraer conocimiento a partir de los datos para apoyar la toma de decisiones es ahora indispensable en cualquier sistema de base de datos, pero los sistemas de bases de datos convencionales están concebidos, diseñados y optimizados para operaciones cotidianas de una organización o empresa, sin

permitir un análisis profundo de los datos. Por lo anterior, existen otras herramientas para extraer información de un conjunto de datos:

- Herramientas OLTP (*On-Line Transaction Processing*), con interfaz gráfica para realizar consultas y reportes sin usar sentencias SQL.

- Herramientas OLAP (*On-Line Analytical Processing*), permite el análisis multidimensional de los datos con diversos criterios de agrupamiento.

- Herramientas de Minería de Datos, DM (*Data Mining*), permiten descubrir patrones, asociaciones, identificar tendencias y comportamientos dinámicos.

Las herramientas de Minería de Datos nacen de la necesidad del Descubrimiento de Conocimiento en Bases de Datos. La Minería de Datos se puede definir como un proceso analítico diseñado para explorar grandes cantidades de datos (generalmente datos de negocio y mercado) con el objetivo de detectar patrones de comportamiento o relaciones entre las diferentes variables. En la Minería de Datos confluyen varias disciplinas, en especial la Estadística, que puede constituirse en un aliado muy productivo y eficaz para los gestores de bases de datos.

El auge informático de la última década ha venido planteando, que para analizar los datos y descubrir relaciones entre sus atributos, es necesario la construcción de una Bodega de Datos. Es importante aclarar, que una Bodega de Datos es una base de datos de información histórica e integrada de distintos sistemas de una empresa para el análisis multidimensional de los datos, enfocada al negocio no a las operaciones.

Las herramientas para el Descubrimiento de Conocimiento en Bases de Datos son generalmente costosas e independientes de las bases de datos operacionales. En los últimos años, los gestores de bases de datos específicamente los gestores de bases de datos comerciales como *ORACLE* y *SQL Server* han incorporado algunos algoritmos o técnicas de Minería de Datos (Berger y Haberstroh, 2005; Dumler, 2005; Larson, 2006; Planeaux *et al.*, 2007; Stackowiak *et al.*, 2007; Utley, 2005).

Los inconvenientes de utilizar las herramientas de Minería de Datos que ofrecen los gestores de bases de datos y las Bodegas de Datos actuales para descubrir conocimiento, se pueden resumir así:

- Requieren de personal altamente calificado que domine la terminología de la Estadística o de la Inteligencia Artificial, manipulen las herramientas informáticas especializadas e interprete los resultados.

- Implican altas inversiones por el uso de herramientas comerciales, dado que las herramientas de distribución libre no tienen la robustez requerida para el manejo de grandes volúmenes de datos o la comunicación con los sistemas gestores de bases de datos.

Los inconvenientes de utilizar paquetes estadísticos para el análisis y el planteamiento de hipótesis en el Descubrimiento de Conocimiento en Bases de Datos, se pueden resumir en:

- Al ser herramientas independientes del almacenamiento de los datos, se requiere de tiempo para la preparación, importación o vinculación de los datos al paquete estadístico, prolongando así el tiempo de respuesta de los análisis y por ende su eficacia.

- Requiere expertos en el área que dominen la terminología estadística, manipulen el paquete y realicen la interpretación de los resultados.

Tradicionalmente, la toma de decisiones se ha basado en juicios altamente subjetivos, pero un nuevo enfoque gerencial toma cada vez más fuerza. A este enfoque gerencial se le denomina *Inteligencia del Negocio* (Business Intelligence o BI), y se basa en la utilización de la información almacenada en bases de datos y en otras fuentes de información internas o externas, para apoyar decisiones con diagnósticos más precisos y soluciones más inteligentes (Soto y Jiménez, 2007).

De la creciente necesidad de Descubrimiento de Conocimiento en Bases de Datos nace en el grupo de Inteligencia Artificial de la Facultad de Minas, Universidad Nacional de Colombia, Sede Medellín, la idea de un mega-proyecto que busque la generación de un conjunto de herramientas gerenciales para fortalecer los sistemas gestores de bases de datos de distribución libre, que facilite la aplicación de un enfoque de Inteligencia de Negocios, con el fin, de promover el Descubrimiento de Conocimiento en Bases de Datos y la interpretación de los resultados en cualquier empresa o institución, sin tener que invertir grandes cantidades de dinero. La presente Tesis de Maestría surge como uno de los primeros proyectos en el mega-proyecto.

Los gestores de bases de datos o herramientas informáticas actuales no son lo suficientemente amigables para que un usuario que no domine la terminología especializada, realice por sí mismo un análisis de regresión multivariante. Sin contar que las herramientas no incorporan la inteligencia suficiente para interpretar los resultados del análisis y presentarlos en forma condensada como nuevo conocimiento, no sugieren, ni recomiendan alternativas según el análisis realizado. Por lo anterior, en la presente Tesis de Maestría se pretende la construcción de los modelos que permitan incorporar las técnicas de regresión lineal y logística multivariante en un gestor de bases de datos, una interfaz para el planteamiento del análisis con la inteligencia suficiente para

facilitar su uso y proporcionar una propuesta de representación gráfica para la presentación de resultados. Adicionalmente, se construye el prototipo de una aplicación Web para validar el modelo propuesto de visualización de resultados y presentar un ejemplo de implementación. Todo con el fin de facilitar el Descubrimiento del Conocimiento en Bases de Datos y por ende, promover la aplicación del enfoque de Inteligencia del Negocio con un sistema que ofrezca los resultados de la manera más amigable para que un directivo pueda tomar las decisiones acertadamente.

El empleo de las técnicas de regresión lineal y logística multivariante tiene como objetivos principales (Neter *et al.*, 1996):

- Explicar la relación entre variables o atributos (obtener un modelo descriptivo de un fenómeno).

- Construir un modelo que permita predecir el valor de la variable respuesta para casos no observados o considerados en un experimento, o la probabilidad de ocurrencia de un suceso.

Algunos beneficios del desarrollo de la presente Tesis de Maestría son los siguientes:

- Ampliar las capacidades de un gestor de bases de datos de distribución libre para el descubrimiento de relaciones entre variables.

- Facilitar el análisis riguroso de la información recopilada en una base de datos y, realizarlo con la misma herramienta, ganando rapidez en los tiempos de respuesta para la toma de decisiones.

- Eliminar la dependencia de especialistas para la manipulación de una herramienta, como paquetes estadísticos que requieran el conocimiento de comandos específicos o dominio de una terminología compleja. Facilitando el trabajo para personas expertas y no expertas en el análisis de datos.

- Incorporar la inteligencia suficiente para interpretar los resultados y facilitar que un usuario final sin conocimientos técnicos, pueda realizar por si mismo un análisis riguroso y con profundidad.

Los productos esperados se pueden resumir en: Un modelo conceptual para la incorporación de las técnicas de regresión lineal y logística multivariante, un modelo físico o funciones programadas en un gestor de bases de datos específico de distribución libre y un prototipo de aplicación mediante una interfaz Web para el planteamiento de análisis de regresión y los resultados del mismo.

El alcance de la presente Tesis de Maestría es la incorporación de la regresión lineal y logística multivariante en un gestor de bases de datos. Para el caso de la regresión logística sólo se considerará la clasificación de objetos en dos clases excluyentes. La propuesta de un modelo para la visualización de resultados y la realización de una aplicación Web como prototipo para garantizar la factibilidad técnica.

En el siguiente capítulo se presentan los fundamentos teóricos, entre los cuales se tiene el concepto de Inteligencia del Negocio, Descubrimiento de Conocimiento en Bases de Datos, enfoques de la Minería de Datos, igualmente se enuncian las técnicas de Minería de Datos presentes en los sistemas gestores de bases de datos; por último, se muestra un ejemplo de análisis con SQL Server. En el capítulo 3, se describen las técnicas de regresión lineal y logística multivariantes, los métodos de solución y sus respectivas pruebas para evaluar la significancia de la regresión y de los coeficientes de regresión. En el capítulo 4, se describe el modelo para la visualización de resultados del análisis de regresión lineal y logística multivariante y, en el capítulo 5, se describe la incorporación de las técnicas de análisis multivariante en un sistema gestor de bases de datos. En el capítulo 6, se describe el prototipo de aplicación Web desarrollada y se realiza la verificación de la factibilidad técnica del modelo propuesto.

Para finalizar, se presentan las conclusiones en el capítulo 7, y se expone brevemente, en el capítulo 8, algunos posibles trabajos futuros que permitirían ampliar las capacidades de un sistema gestor de bases de datos para el Descubrimiento de Conocimiento.

2. FUNDAMENTOS TEÓRICOS

2.1. Inteligencia del Negocio

Algunas definiciones de Inteligencia del Negocio son:

- "La Inteligencia del Negocio es la entrega de información precisa y útil para las personas adecuadas que toman las decisiones dentro de un marco de tiempo justo, para dar soporte a toma de decisiones efectivas" (Larson, 2006).

- "La Inteligencia del Negocio puede ser definida como tener acceso apropiado a los datos o información necesaria para tomar las decisiones apropiadas del negocio en el tiempo apropiado" (Stackowiak *et al.*, 2007).

La Inteligencia del Negocio es un nuevo enfoque gerencial que se basa en la utilización de la información almacenada en bases de datos y en otras fuentes de información internas o externas, para apoyar la toma de decisiones. La Tabla 1, compara el enfoque tradicional para la toma de decisiones con el enfoque de Inteligencia del Negocio.

Tabla 1. Paralelo entre enfoques gerenciales

Tradicional	Inteligencia del Negocio
Basado en juicios subjetivos	Basado en juicios objetivos
Basado en la intuición y en las emociones	Basado en la información
Utiliza poca información	Aprovecha la información disponible
	La información apoya la toma de decisiones

En la actualidad las organizaciones se enfrentan a un mundo cada vez más competitivo y, por tanto, las estrategias deben ser flexibles para adaptarse a las condiciones cambiantes del entorno. El análisis de información, basado en el enfoque de Inteligencia del Negocio podría permitir conocer tendencias en variables de interés, sus relaciones o asociaciones, evaluar el impacto de decisiones tomadas en el pasado, entre otras. Por esto, es necesario facilitar el análisis de los datos para la toma de decisiones más acertadas y oportunas.

La iniciativa de un proyecto de Descubrimiento de Conocimiento en Bases de Datos debe estar dirigida al negocio y no a la tecnología. Es decir, la justificación de la iniciativa no debe ser sólo el hecho de experimentar con nuevas tecnologías, sino reducir los problemas que afectan la rentabilidad o la eficiencia de la organización, teniendo en cuenta, la importancia de la calidad

de la información que constituye la diferencia entre tomar decisiones correctas e incorrectas.

El enfoque de Inteligencia del Negocio busca maximizar la explotación de la información recolectada por una organización, al ser común la práctica de recoger más información de la que en realidad usan. También es común pensar que la Minería de Datos requiere grandes inversiones y costosas aplicaciones informáticas, que sólo son justificables en grandes empresas, pero es sorprendente el poco aprovechamiento de los recursos informáticos disponibles. Un desafío es hacer que la extracción de la información y los análisis se conviertan en una operación rutinaria y semiautomática.

2.1.1. Dificultades para la implementación

Las dificultades para la implementación del enfoque de Inteligencia del Negocio en cada organización son diferentes, y dependerá de un sin número de factores como: Tipo de organización, dinámica del mercado específico, cantidad de personal, formación académica del personal, entre otros. En general, se pueden resaltar las siguientes dificultades:

Inercia organizacional
La inercia organizacional hace referencia a la resistencia a los cambios que se presentan en las organizaciones, dado que todo el personal no asimila los cambios con la misma rapidez.

Personal poco capacitado
Es frecuente que las organizaciones tengan un bajo número de personal capacitado para el análisis de información, y poco personal actualizado en la utilización de herramientas informáticas. El personal operativo es necesario en toda organización para su funcionamiento, pero el mundo cada vez más competitivo, obliga a las organizaciones a contratar personal altamente calificado.

Brecha tecnológica
La reducción de gastos en las organizaciones frecuentemente converge a un bajo nivel de actualización tecnológica.

Dificultades en el análisis de datos
En el empleo de paquetes estadísticos o programas especializados se requiere de tiempo y esfuerzo en la preparación de los datos, su importación o vínculo. Adicionalmente, se pone a la vista la dependencia de expertos que dominen las técnicas y la terminología, manipulen las aplicaciones e interpreten los resultados de los análisis. Por todo lo anterior, se mantiene una baja eficiencia en el proceso completo de descubrimiento de conocimiento.

2.2. Descubrimiento de Conocimiento en Bases de Datos

Los sistemas de consulta-respuesta basados en SQL pueden responder de manera espontánea a una gran cantidad de preguntas, algunas de ellas de bastante complejidad que representan parte del conocimiento almacenado en una base de datos. Sin embargo, existe otro conocimiento oculto, representado en gran parte por relaciones entre distintos objetos de la base de datos, que demanda un análisis mucho más complejo de los datos almacenados.

Los datos almacenados en una base de datos pueden contener información que no se ve a simple vista, el proceso completo de extracción de conocimiento implícito en los datos, es denominado Descubrimiento de Conocimiento en Bases de Datos o KDD, por sus siglas en inglés “Knowledge Discovery in Databases” (Han y Kamber, 2001).

En los últimos años, el Descubrimiento de Conocimiento en Bases de Datos, ha recibido especial atención debido a la disponibilidad actual de grandes cantidades de datos y a la necesidad de convertirlos en información útil y en nuevo conocimiento (Jiménez, 2008). Este proceso también es conocido como Minería de Datos: un proceso que a través del análisis y la cuantificación de relaciones en los datos, permite extraer patrones comunes, asociaciones o modelar el comportamiento de distintos fenómenos de la naturaleza. Siendo rigurosos, la Minería de Datos es el proceso central del Descubrimiento de Conocimiento en Bases de Datos, pero se necesitan otros procesos antes de aplicar los métodos de la Inteligencia Artificial, de la Estadística o de las Bases de Datos.

2.2.1. Pasos del Descubrimiento de Conocimiento

Implantar un proceso para transformar datos en información, información en conocimiento y conocimiento en ayuda a la toma de decisiones, no es tan difícil ni complejo como se tiende a pensar; máxime si se tienen claros los objetivos y las necesidades. Los atributos que se buscan son siempre los mismos: la información debe ser tangible, precisa, comprensible y oportuna.

A continuación se detallan los pasos en el proceso de Descubrimiento de Conocimiento en Bases de Datos (Jiménez, 2008; Mitra y Acharya, 2003; Soto y Jiménez, 2007).

Comprensión del dominio de aplicación
Incluye la recolección de la información a priori relevante, sobre la temática que se aborda en el dominio del problema y de los supuestos que se cumplen. También es común dividir esta etapa en dos:

a) *Comprensión del negocio*: Esta fase inicial se enfoca en la comprensión de los objetivos del proyecto y exigencias desde una perspectiva de

negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de Minería de Datos y en un plan preliminar diseñado para alcanzar los objetivos. Al comienzo de un proceso de Descubrimiento de Conocimiento en Bases de Datos, el usuario a menudo no conoce ni el objetivo preciso del análisis ni la naturaleza exacta de los datos.

b) *Comprensión de los datos*: La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que le permiten familiarizarse primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

La exploración inicial del análisis de datos puede ayudar a los usuarios a entender la naturaleza de los datos y formar hipótesis potenciales de la información oculta. La estadística descriptiva simple y las técnicas de visualización, proporcionan las primeras ideas sobre los datos. Por ejemplo: La distribución de clientes por edad y regiones geográficas pueden dirigir futuras estrategias de comercialización.

Extracción, transformación y carga de datos

Se eligen los datos que se consideran relevantes para el análisis y se hace un proceso de integración de datos si están almacenados en distintas fuentes o con distintos formatos. La extracción de datos se enfocan en el problema y debe ser consistente con los objetivos del proyecto definidos en la etapa anterior.

A veces, una recopilación y resumen de los datos sólo puede ser un objetivo de un proyecto de Descubrimiento de Conocimiento en Bases de Datos. Esta clase de problema estaría en lo más bajo de la escala de problemas en Descubrimiento de Conocimiento en Bases de Datos. Sin embargo, la recopilación y resumen de datos ocurren generalmente en combinación con otros tipos de problemas de Minería de Datos.

Preprocesamiento de datos

Este proceso se necesita para depurar la información, chequear inconsistencias o preparar los datos para la minería. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces, y en ocasiones, en un orden prescripto. El preprocesamiento puede incluir tareas de:

a) *Limpieza de datos*: Consiste en la remoción de ruido (errores) o corrección de datos inconsistentes.

b) *Transformación de los datos*: En ocasiones, los datos deben ser transformados o consolidados en una forma apropiada para la minería.

Puede ser necesario resumir la información recolectada, realizar cambios de escala o reducir la dimensionalidad del problema, antes de aplicar una técnica de Minería de Datos, en particular.

Minería de Datos

La Minería de Datos se define como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de los datos (Frawley *et al.*, 1992). Esta fase es el proceso central del Descubrimiento de Conocimiento en Bases de Datos donde se aplican los métodos o técnicas que permiten el análisis de los datos para encontrar relaciones implícitas o patrones previamente desconocidos. La Minería de Datos consta de un amplio espectro de técnicas para la caracterización de un dominio, la discriminación o la clasificación de objetos, o para el hallazgo de asociaciones o dependencias funcionales, entre otras tareas.

En esta fase varias técnicas de modelado son seleccionadas, aplicadas, y sus parámetros son calibrados a valores óptimos. En general, hay varias técnicas para el mismo tipo de problema de Minería de Datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, es a menudo necesario volver a la fase de preparación de datos.

La Minería de Datos invierte la dinámica del método científico, dado que se coleccionan los datos y se esperan que de ellos emerjan hipótesis, mientras que el método científico, formula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis.

Evaluación de los hallazgos

Este proceso es necesario para validar los modelos construidos o las hipótesis planteadas. En la validación de modelos, es frecuente determinar el grado de bondad de ajuste de datos reales a los modelos, empleando datos de prueba o validación.

Antes de proceder al despliegue final del modelo, también se debe realizar una revisión de los pasos ejecutados para comparar el modelo correctamente obtenido con los objetivos del negocio. Un objetivo clave es determinar si existe algún tema importante del negocio que no ha sido suficientemente considerado.

Es frecuente no poder reconocer la división entre la evaluación y la interpretación de los hallazgos, dado que validar o comprobar que las conclusiones son válidas y satisfactorias, incluye interpretación. En caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar e interpretar los modelos en busca de aquel que mejor se ajuste al dominio de aplicación. Si ninguno de los modelos alcanza los resultados esperados, se regresará a una etapa anterior en busca de nuevos modelos.

Interpretación y presentación del conocimiento

La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado de modo que el analista o las personas que toma decisiones puedan usarlo, lo cual requiere la aplicación de técnicas para la visualización y la representación del conocimiento minado. Al final de esta fase, una decisión en el uso de los resultados de Minería de Datos debería ser obtenida dentro de un proceso de toma de decisiones de una organización.

El proceso de Descubrimiento de Conocimiento en Bases de Datos aparenta ser un proceso en cascada, pero realmente es un proceso cíclico, y desde cualquier etapa se puede regresar a otra, la retroalimentación incluso alcanza a la base de datos, dado que puede servir de guía para cambiar, aumentar o disminuir la información almacenada. En la Figura 1, se ilustra el proceso de Descubrimiento de Conocimiento en Bases de Datos.

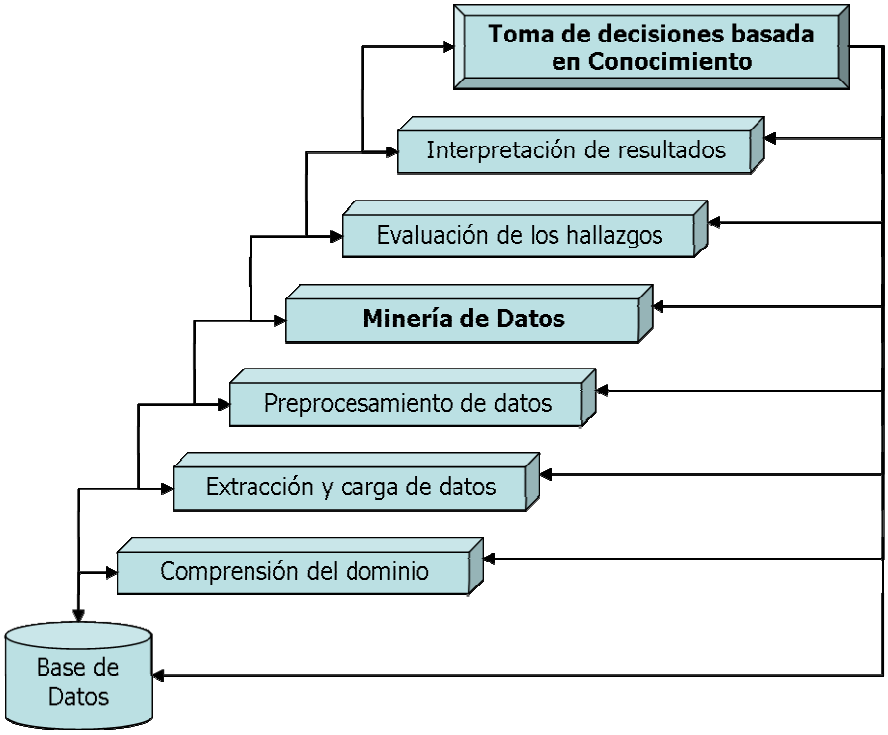


Figura 1. Proceso Descubrimiento de Conocimiento en Bases de Datos.

Esta Tesis de Maestría aporta principalmente en los tres últimos pasos del proceso de Descubrimiento de Conocimiento en Bases de Datos. En la etapa de *Minería de Datos* se ofrecen las técnicas de regresión lineal y logística multivariantes en un sistema gestor de bases de datos, adicionalmente, en el prototipo de aplicación Web se facilita la selección de datos para el planteamiento del modelo de regresión, sin necesidad de que el usuario posea conocimientos en SQL.

Los procedimientos incorporados en un sistema gestor de bases de datos cubren la etapa de *Evaluación de los hallazgos* con la validación de los modelos de regresión construidos, por medio de pruebas de hipótesis para hallar la significancia de la regresión, de los coeficientes de regresión y la validación de los supuestos del modelo. Por último, el aporte en la etapa de *Interpretación y presentación del conocimiento* corresponde a un modelo propuesto para la visualización de los resultados del análisis.

2.2.2. Algunos enfoques en Minería de Datos

En definitiva, la Minería de Datos es una tecnología usada para descubrir información oculta, desconocida y potencialmente útil.

A continuación se describen brevemente algunos enfoques en Minería de Datos.

Segmentación y/o clasificación

En la literatura existe ambigüedad en el significado de los términos *segmentación*, *agrupamiento* y *clasificación*, y en ocasiones se utilizan en forma indiferente. A veces se refieren a la creación de grupos, creación de clases o creación de modelos para predecir las clases conocidas para casos antes no vistos.

La segmentación apunta a la separación de los datos en subgrupos o clases significativos, todos los miembros de un subgrupo deben compartir características comunes (Chapman *et al.*, 2000). El analista puede suponer ciertos subgrupos como relevantes basado en un conocimiento previo de los datos o como resultado de la descripción y el resumen de datos. Adicionalmente, existen técnicas automáticas de agrupamiento (clustering) que pueden descubrir las estructuras antes insospechadas y ocultas en que permiten la segmentación. La segmentación a veces puede ser un objetivo de la Minería de Datos.

A menudo, la segmentación es un paso hacia la solución de otros tipos de problemas. Por lo tanto, el objetivo es encontrar los subconjuntos de datos homogéneos que son más fáciles para analizar. Típicamente en grandes conjuntos de datos variados, se obscurecen patrones o relaciones importantes. En este caso, la segmentación apropiada hace la tarea más fácil y/o más significativa. Por ejemplo: Una empresa de venta de autos con regularidad recoge información sobre sus clientes acerca de sus características socioeconómicas como el ingreso, la edad, el sexo, la profesión, ingresos promedios, entre otros. Realizando una segmentación, la empresa puede dividir a sus clientes en subgrupos más comprensibles y analizar la estructura de cada subgrupo, para definir o desarrollar estrategias de comercialización específicas para cada grupo separado.

La clasificación asume que hay un conjunto de objetos caracterizados por algún atributo o rasgo, que pertenecen a diferentes clases. La etiqueta de clase es un valor discreto (simbólico) y es conocido para cada objeto. El objetivo es construir los modelos de clasificación (a veces llamados clasificadores), que asignan la etiqueta de clase correcta a objetos antes no vistos y sin etiquetas. En general, los modelos de clasificación se pueden utilizar como modelos predictivos (Chapman *et al.*, 2000).

Las etiquetas de clase pueden ser definidas por el usuario o derivadas de la segmentación. La clasificación es uno de los tipos de problemas más importantes de la Minería de Datos que ocurren en una amplia gama de aplicaciones. Muchos problemas de la Minería de Datos pueden ser transformados a problemas de clasificación. Por ejemplo: El problema de evaluar el riego potencial de otorgar un crédito a un cliente nuevo, puede ser transformado a un problema de clasificación donde se crean dos clases (clientes buenos y clientes malos). Un modelo de clasificación puede ser generado de los datos de comportamiento crediticio de los clientes existentes en una base de datos. Finalmente, el modelo de clasificación puede ser usado para determinar a cual de a una las dos clases pertenece el cliente nuevo y por ende decidir el otorgamiento del crédito.

La clasificación tiene conexiones a casi todos los otros tipos de problemas. Los problemas de predicción pueden ser transformados a los problemas de clasificación por discretización de etiquetas de clase continuas, porque las técnicas de discretización permiten transformar rangos continuos en intervalos discretos. Estos intervalos discretos, más que los valores numéricos exactos, son usados como etiquetas de clase, y de ahí conducen a un problema de clasificación. Algunas técnicas de clasificación producen una clase comprensible o descripciones de concepto. Hay también una conexión al análisis de dependencia porque los modelos de clasificación típicamente usan y aclaran las dependencias entre atributos.

La clasificación requiere de categorías que puedan reunir un grupo de observaciones y que se distinga de otro grupo (Hand, 1989). La clasificación no es única y un agrupamiento puede ser bueno para ciertas cosas e inadecuado para otras. El análisis de grupo es principalmente una herramienta para la exploración de datos, por lo cual se tiene una clasificación desconocida (Hand, 1989). Un análisis de grupo puede generar una clasificación, para esto se requieren dos cosas:

- Una medida que muestre cómo los subespacios de representación se ajustan aproximadamente a la representación original. Esta medida se basa en la similitud.
- Un algoritmo para buscar que los subespacios se optimicen.

Los algoritmos se clasifican en: Métodos jerárquicos y métodos de optimización. Los métodos de análisis de grupo jerárquico se dividen a su vez en métodos de acumulación o división.

Métodos jerárquicos de acumulación: Se parte de *n* grupos, iguales a *n* observaciones. Un criterio de mínima distancia o similaridad permite ir formando un grupo con las dos observaciones más cercanas, hasta que todas las observaciones pertenecen a un único grupo y se puede representar en un gráfico de árbol o dendrograma.

Métodos jerárquicos de división: Inicialmente el conjunto de observación forma un solo grupo o grupo padre, se inicia la división formando dos subgrupos y se continúa hasta poder obtener un gráfico de árbol o dendrograma. La decisión de dividir un grupo en dos subgrupos, debe considerar todas las variables simultáneamente, dependiendo de la técnica seleccionada.

Métodos de optimización: Estos métodos requieren definir el criterio de agrupamiento y un método de optimización. Cada criterio de optimización da lugar a una estructura de grupos, por ello se sugiere explorar los datos con distintos métodos. Los métodos de optimización consisten en transferir puntos u observaciones entre grupos, buscando la optimización.

Los anteriores métodos de clasificación son técnicas de agrupamiento que deben cumplir con los requisitos de cobertura y exclusividad. La cobertura hace referencia a que la unión de todos los grupos forma el universo del discurso; la exclusividad hace referencia a la pertenencia única de un elemento a un grupo.

La segmentación puede también proporcionar las etiquetas de clase o restringir el conjunto de datos para que buenos modelos de clasificación puedan ser construidos. Es útil analizar desviaciones antes de que un modelo de clasificación sea construido. Las desviaciones y contingencias (los valores atípicos - outliers) pueden obscurecer el patrón que podría permitir un buen modelo de clasificación. De otro modo, un modelo de clasificación también puede ser usado para identificar desviaciones y otros problemas con los datos.

Predicción

Otro tipo de problema importante que ocurre en una amplia gama de usos de las técnicas de minería de datos es la predicción. La predicción es similar a la clasificación, se diferencia en que la predicción del atributo objetivo es continuo y no un atributo cualitativo discreto o clase. El objetivo de la predicción busca encontrar el valor numérico del atributo objetivo para objetos no vistos. En la literatura, este tipo de problema es comúnmente llamado regresión. Si la predicción trata con datos de serie de tiempo, entonces se le llama pronosticación.

El análisis de regresión es una técnica estadística para la generación de modelos y la investigación de relaciones entre variables que sean cuantitativas, cualitativas, o de ambos tipos (Pérez, 2004). Es común que existan relaciones entre dos o más variables, pero cuando estas relaciones no han sido identificadas o no están completamente determinadas, toma valor el análisis de regresión. En los últimos años (gracias en gran parte al desarrollo de los computadores), el análisis de regresión y en especial los métodos de análisis multivariantes han probado su eficacia en diferentes áreas: las ingenierías, las ciencias de la salud, en diversas investigaciones, entre otras. Los métodos de análisis multivariantes de datos han probado su eficacia, sobre todo en los métodos factoriales y de clasificación, muy utilizados en la Minería de Datos (Pérez, 2004). En el capítulo 3, se describe en detalle las técnicas de regresión lineal y logística multivariante con sus criterios de interpretación.

Análisis de dependencia

El análisis de dependencia consiste en encontrar un modelo que describe dependencias significativas (o asociaciones) entre atributos o acontecimientos. Aunque las dependencias pueden ser usadas para el modelado predictivo son más usados por su comprensión. Las dependencias pueden ser estrictas o probabilísticas (Chapman *et al.*, 2000).

La asociación es un caso especial de dependencia. Las asociaciones describen las afinidades de atributos, esto es, atributos o acontecimientos que con frecuencia ocurren simultáneamente (Chapman *et al.*, 2000).

Los algoritmos para detectar asociaciones son muy rápidos y producen muchas asociaciones. Seleccionar las más importantes es un desafío.

El análisis de dependencia tiene conexiones cercanas a la predicción y a la clasificación, ya que las dependencias implícitamente son usadas para la formulación de modelos predictivos.

En aplicaciones, el análisis de dependencia a menudo converge a la segmentación. Cuándo las dependencias no son significativas en grandes conjuntos de datos es aconsejable realizar el análisis sobre segmentos de datos más homogéneos.

2.3. Gestores de bases de datos con enfoque a la Inteligencia del Negocio

En los últimos años, los sistemas gestores de bases de datos comerciales como ORACLE y SQL Server, ofrecen herramientas para la construcción de Bodegas de Datos, la utilización de la tecnología OLAP y algunos algoritmos para Minería de Datos.

Las Bodegas de Datos hacen parte de los sistemas de apoyo a la toma de decisiones, que guardan información histórica resumida, consolidada y usualmente, requieren información de muchas fuentes, incluso de fuentes de información externas a la organización. La utilización de las Bodegas de Datos y la tecnología OLAP, introducen el concepto de Cubo de Datos, también denominado como Cubo Multidimensional o simplemente Cubo (Zvenger y Fidel, 2005).

El Cubo de Datos constituye el modelo de datos de una base de datos multidimensional; en un cubo, la información se representa por medio de matrices multidimensionales o cuadros de múltiples entradas, que permite realizar distintas combinaciones de sus elementos para visualizar los resultados desde distintas perspectivas y variando los niveles de detalle. El diseño de un cubo exige determinar qué se quiere capturar como medida, es decir, los valores cuantitativos que se quieren analizar y monitorear como indicadores de la actividad del negocio.

Las herramientas OLAP permiten el análisis multidimensional interactivo de los datos con diversos criterios de agrupamiento, esta exploración interactiva distingue a OLAP de las herramientas simples de consulta y reportes. El rol de OLAP va mas allá del monitoreo de medidas de ejecución. La definición de tales medidas debería ser un proceso de articulación de valores y metas. En muchos casos, este proceso es tan importante como los resultados. Las herramientas OLAP pueden proveer un entorno colaborativo y son importantes por la multidimensionalidad, que permite ver las medidas del negocio desde varias perspectivas, trabajar en forma intuitiva con datos agregados o totalizados. Además, convierte al usuario en un explorador activo de la información, permitiéndole ejecutar consultas complejas que involucran muchas facetas de su negocio sin usar sentencias SQL.

2.3.1. Modelamiento y Clasificación de Herramientas OLAP

Las Bodegas de Datos pueden requerir grandes espacios de almacenamiento, y por su tamaño, la eficiencia de los métodos de acceso y las técnicas de procesamiento de consultas suelen ser parámetros importantes en el diseño de bases de datos multidimensionales. Las características relevantes en los servidores de las Bodegas de Datos son: El manejo de los índices, la materialización de las vistas, las técnicas para solucionar consultas complejas y el paralelismo en el procesamiento masivo de bases de datos. Un elemento esencial de la arquitectura de las Bodegas de Datos es la administración de los metadatos. Entre los diferentes tipos de metadatos a administrar se encuentran: Los metadatos de administración, negocio y operaciones (Chaudhuri y Dayal, 1997).

El modelo conceptual de los sistemas de soporte de decisión que influencia el diseño de las Bodegas de Datos es un modelo multidimensional de datos, que

facilita el análisis y visualización de los datos, en el cual, los diagramas de entidad-relación y las técnicas de normalización usadas generalmente en bases de datos operacionales no son apropiados dado que compromete la eficiencia de las consultas. Las bases de datos relacionales están optimizadas para obtener una ejecución optima en consultas simples y frecuentes, pero no funcionan de manera ideal para las consultas multidimensionales y complejas, ya que existen muchas de ellas que no se pueden expresar en una única consulta SQL, y seguramente se requerirán muchas operaciones de JOIN, lo cual reduce drásticamente el tiempo de respuesta de la consulta.

Existen tres tipos de estrategias de almacenamiento de información en las herramientas OLAP (Zvenger y Fidel, 2005):

MOLAP

Las bases de datos multidimensionales especializadas usan estructuras de tipo arreglo para almacenar los datos. Estas estructuras están indexadas con el fin de proveer un tiempo de acceso óptimo a cualquier elemento. Se pueden distinguir dos enfoques en la forma de organizar estas estructuras: *Arquitectura de Hipercubo y Arquitectura de Multicubos*.

Arquitectura de Hipercubo: Almacena un único gran cubo en el cual cada medida está referenciada y totalizada en todas las dimensiones del mismo, con una ejecución más pareja en cuanto al tiempo de respuesta a las consultas; pero requiere mucho espacio en disco, y además necesita un buen manejo de la dispersión de los datos para evitar que el tamaño del cubo se vuelva inmanejable.

Arquitectura de Multicubos: Los datos se guardan en más de un cubo, y se logra una mejor ejecución si la consulta no requiere el acceso a más de un cubo, pero en el caso contrario, la ejecución se reduce notoriamente ya que se requiere un procesamiento complejo para asegurar que los resultados del cruce de cubos sea consistente.

ROLAP

Las Bodegas de Datos ROLAP son construidas sobre una tecnología relacional, pero la optimización se dirige al apoyo de toma de decisiones en lugar de las operaciones transaccionales. La arquitectura MOLAP presenta una mejor ejecución para el análisis multidimensional, pero la arquitectura ROLAP tiene ventajas en otros aspectos. En particular, son escalables a conjuntos más grandes de datos e incluyen soporte para replicación, “rollback” y recuperación. Además, en la mayoría de las organizaciones están más familiarizadas con el uso de una base de datos relacional.

Las herramientas ROLAP brindan análisis multidimensional sobre datos almacenados en una base de datos relacional. Lo hacen a través de un mapeo entre los datos en la bodega a un modelo multidimensional, usando consultas

SQL. Para mejorar la ejecución, se tiende a almacenar algunos valores totalizados en la bodega, así que los datos dispersos siguen siendo un tema de importancia que en este caso se delega al diseñador. Cuando las Bodegas de Datos llegan al orden del terabyte, se observa claramente ventajas con respecto a la arquitectura MOLAP.

HOLAP

Las herramientas con arquitectura HOLAP incluyen características de ambos modelos, MOLAP y ROLAP. Cada alternativa tiene sus ventajas y desventajas. En lugar de discutir cual de las dos es mejor, se debe definir un criterio para optar por una u otra, y evaluar el alcance de HOLAP, que en la práctica, intenta combinar lo mejor de ambos modelos.

2.3.2. Técnicas de Minería de Datos incorporados en gestores de bases de datos comerciales

Las últimas versiones de los gestores de bases de datos ORACLE y SQL Server, han incorporado algunos algoritmos o técnicas para el análisis de datos, buscando facilitar la implementación del enfoque de Inteligencia del Negocio.

ORACLE Data Mining

La herramienta de Minería de Datos del gestor de bases de datos ORACLE, no es una herramienta independiente, funciona en conjunto con el motor de base de datos. A continuación se enuncian las técnicas que ofrece (Planeaux *et al.*, 2007; Berger y Haberstroh, 2005):

Clasificador Bayesiano "Naïve" (Naive Bayes): Técnica de clasificación y predicción que construye modelos que predice la probabilidad de posibles resultados. Naive Bayes utiliza datos históricos para encontrar asociaciones y relaciones y hacer predicciones. Este algoritmo predice resultados binarios o multiclase.

Árboles de Decisión: Técnica basada en los algoritmos de árboles de regresión y clasificación, en cada nodo se tiene un criterio de separación. Los árboles de decisiones son populares porque son universalmente aplicables, fáciles de entender y aplicar.

Máquinas de Vector de Soporte (Support Vector Machines): El algoritmo soporta modelos de clasificaciones binarias y multiclase, predicción y regresión. Este algoritmo es particularmente bueno para descubrir patrones ocultos en los problemas que tienen un número muy grande de atributos independientes, con un número muy limitado de registros o de observaciones.

Atributo Relevante: El algoritmo permite identificar los atributos de mayor influencia sobre un atributo respuesta.

Agrupamiento: ORACLE provee dos algoritmos, K-Means Realzado y Clustering Ortogonal (O-Cluster). Las técnicas permiten identificar grupos en una población de datos.

Reglas de Asociación: Las reglas de asociación detectan eventos asociados que se ocultan en las bases de datos. Las reglas de asociación generan un conjunto de pares A-B con una confianza y un soporte determinado.

Característica de Selección: El algoritmo Nonnegative Matrix Factorization (NMF) es útil para reducir un gran conjunto de atributos en atributos representativos, similar en su concepto al Análisis de Componentes Principales, pero capaz de manipular cantidades mayores de atributos y, en un modelo aditivo de representación, NMF es un algoritmo poderoso de Minería de Datos de avanzada tecnología que puede servir en una variedad de casos de uso.

Detección de Anomalías: El algoritmo permite la detección de "casos raros", aún con muy pocos ejemplos disponibles. ORACLE puede “clasificar” los datos en "normal" y "anormal", el algoritmo usa una versión del algoritmo “Support Vector Machines” para crear un perfil de una clase conocida. Cuando el modelo se aplique a la población general, los casos clasificados como anormales son datos “sospechosos”.

Minería de Texto: Permite realizar Minería de Datos en datos no estructurados, como es el caso del texto.

Las primeras cuatro técnicas enunciadas son técnicas de Minería de Datos supervisadas y requieren de la interacción con el usuario, las demás son técnicas no supervisadas.

SQL Server

El Gestor de bases de datos SQL Server incorpora una herramienta que ofrece los siguientes algoritmos o técnicas de Minería de Datos (Utley, 2005; Dumler, 2005):

Árboles de Decisión

Reglas de Asociación

Clasificador Bayesiano “Naïve” (Naive Bayes)

Agrupamiento (Sequence Clustering)

Minería de Texto

Series de Tiempo: El algoritmo de Series de Tiempo se usa para analizar y pronosticar datos basados en tiempos. Este algoritmo busca patrones a través de la serie de datos múltiple.

Redes Neuronales: El algoritmo busca revelar relaciones en los datos que otros algoritmos omiten, mientras el algoritmo de Redes Neuronales tiende a ser más lento que los otros algoritmos y encuentra relaciones que pueden ser poco intuitivas.

Las técnicas que no presentan ninguna descripción son similares a las presentadas por ORACLE.

SQL Server incluye la técnica de regresión lineal como un caso particular del algoritmo de *Árboles de Decisión*, de forma similar, incluye la técnica de regresión logística como un caso particular del algoritmo de *Redes Neuronales*.

2.3.3. Ejemplo de análisis con SQL Server

La herramienta de SQL Server para realizar análisis de los datos es "SQL Server Business Intelligence Development Studio", esta herramienta se especializa en la creación, mantenimiento y despliegue de proyectos específicos de Inteligencia del Negocio. Se destaca la posibilidad de gestionar los proyectos al estilo de los proyectos de desarrollo de *Visual Studio.*

Se utilizará la base de datos "benchmark", que contiene información sobre 398 automóviles y es proporcionada por el Laboratorio de Tecnología de la Información (ITL) del Instituto Nacional de Estándares y Tecnología (NIST) del gobierno de Estados Unidos, con el propósito de permitir a los investigadores analizar el comportamiento de diversas técnicas estadísticas o de aprendizaje de máquinas (ITL, 2006). La base de datos contiene información de: *Rendimiento (millas por galón), Número de Cilindros, Desplazamiento del pistón, Potencia, Peso, Aceleración, Modelo y Marca*. Para este ejemplo con regresión lineal se define *Rendimiento* como la variable respuesta, las variables independientes o explicativas son la *Potencia*, el *Peso* y la *Aceleración*. Los resultados que arroja el análisis con esta herramienta se presentan en la Figura 2 y se resalta con un círculo rojo el lugar donde aparecen los coeficientes de regresión.

El ejemplo de regresión logística se realiza con los datos de la tabla 14.3 del libro de Neter *et al.* (1996), los datos están disponibles en el CD-ROM anexo al libro. La variable respuesta es la variable indicadora binaria *Enfermedad*. Las variables explicativas son: *Edad, Socioeconómico 1, Socioeconómico 2* y *Sector*. Los pasos necesarios para el análisis de regresión logística son similares a los enunciados para la regresión lineal, la diferencia se presenta en el momento de definir el modelo de Minería de Datos. La Figura 3, presenta los resultados del análisis en SQL Server.

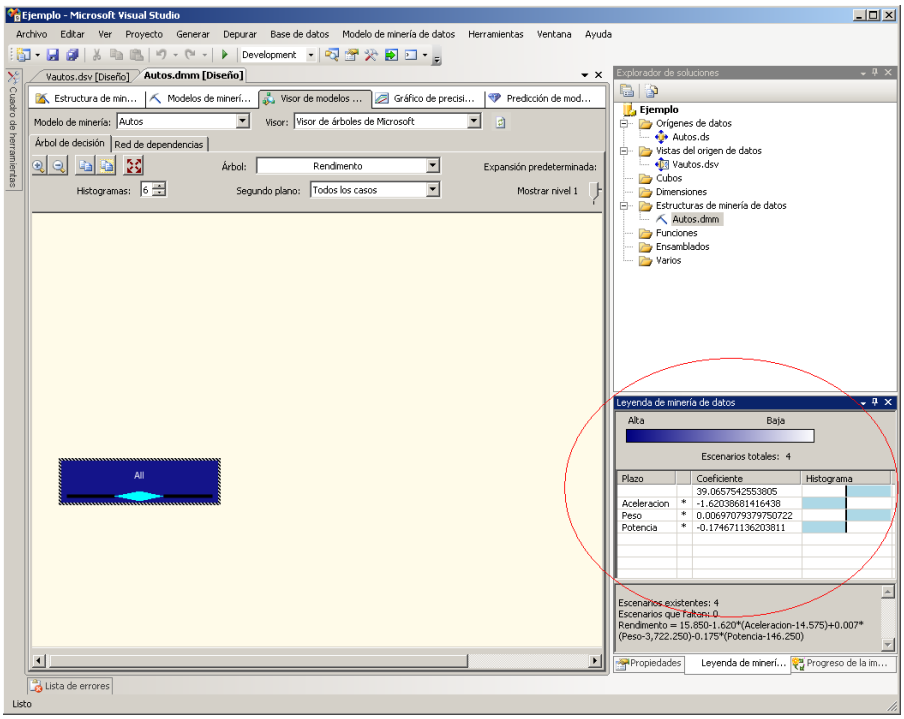


Figura 2. Resultados con SQL Server para el análisis de regresión lineal

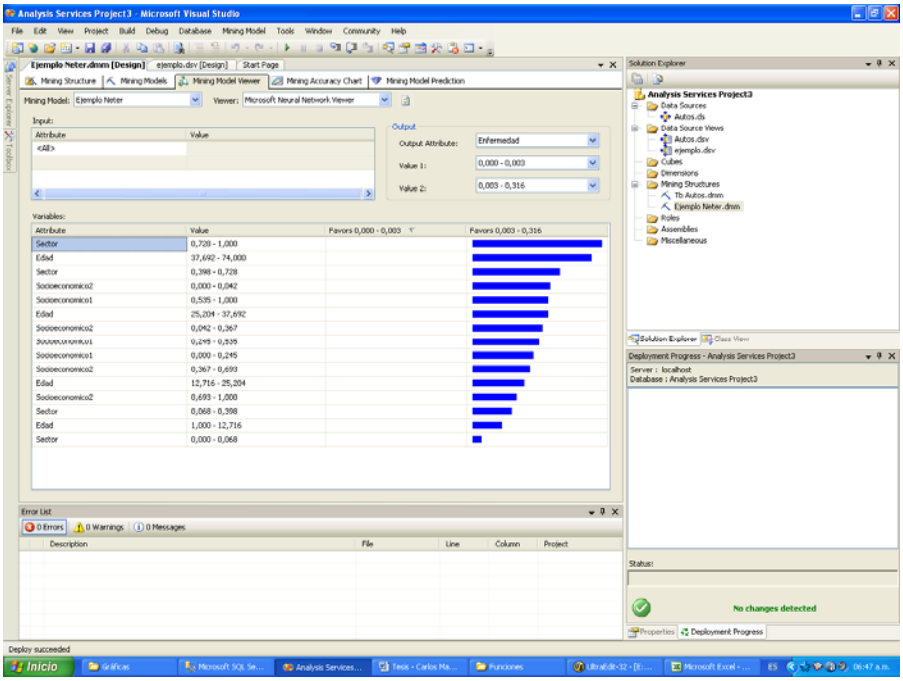


Figura 3. Resultados con SQL Server para el análisis de regresión logística

En Larson (2006) o Watt (2006), se encuentra un mayor detalle de los pasos o procedimientos para realizar análisis con Microsoft SQL Server.

SQL Server ofrece las técnicas que se proponen incorporar en la presente Tesis de Maestría, pero con algunas limitaciones como:

- Requieren de personal altamente calificado para que manipule las herramientas, aunque ofrece asistentes en la mayoría de los pasos, la herramienta no es intuitiva y confunde a usuarios no informáticos. La plataforma con características similares a la plataforma de proyectos de desarrollo de *Visual Studio* es ideal para programadores pero inapropiado para los usuarios que aquí se contemplan.

- La presentación de resultados no es rigurosa ni clara, en el caso de la regresión logística, no presenta los coeficientes, ni el modelo matemático esperado en este análisis. En ningún caso, muestra detalles del análisis de varianza o de pruebas de significancia y mucho menos gráficos para el análisis de los residuales.

- No es una herramienta de distribución libre.

3. TÉCNICAS MULTIVARIANTES Y CRITERIOS DE VALIDACIÓN

El campo de la Estadística tiene que ver con la recopilación, presentación, análisis y uso de datos para tomar decisiones y resolver problemas (Montgomery y Runger, 2003). En el marco del análisis estadístico de datos se han desarrollado técnicas univariadas y multivariadas, que permiten estudiar una colección de datos. Las técnicas estadísticas para el análisis, también se consideran parte de la Minería de Datos, dado que su propósito es el mismo.

El análisis de regresión es una técnica estadística para el modelado y la investigación de la relación entre dos o más variables (Montgomery y Runger, 2003). En cualquier área es común que existan relaciones entre dos o más variables, pero cuando estas relaciones no han sido identificadas o no están completamente determinadas, toma valor el análisis de regresión. Según Neter *et al.* (1996), el análisis de regresión sirve principalmente para tres propósitos: descripción, control y predicción.

El análisis de regresión multivariante está basado en los siguientes supuestos (Neter *et al.*, 1996):

- Las variables independientes o explicativas deben ser linealmente independientes. Es decir, no debe ser posible que una variable independiente sea explicada por una combinación lineal de las otras.

- Los términos de error deben distribuirse normalmente, con media cero, varianza constante y ser independientes entre sí $N\left(0,\sigma^2\right)$.

Estos supuestos no son difíciles de cumplir y se verifican mediante el análisis de residuales, la detección de valores atípicos o influyentes y las pruebas de independencia. Además, cuando los supuestos no se cumplen es posible aplicar medidas correctivas en la mayoría de los casos.

En los últimos años, el análisis de regresión, los métodos factoriales y los métodos de clasificación, han probado su eficacia en diferentes áreas: Ingenierías, ciencias de la salud e investigaciones, entre muchas otras. (Pérez, 2004).

La presente Tesis de Maestría centra su atención en la regresión lineal y logística multivariante, que se detallan a continuación.

3.1. Regresión lineal multivariante

La regresión lineal multivariante o múltiple es un modelo de regresión que considera $p-1$ regresores o variables independientes para explicar una variable de respuesta o variable dependiente Y. La forma general del modelo de regresión lineal multivariante, es (Weisberg, 2005):

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (1)$$

Sean $\beta_0, \beta_1, \beta_2, ..., \beta_{p-1}$ los parámetros del modelo que representa los coeficientes de regresión.

Sea $X_{i,1}, X_{i,2}, ..., X_{i,p-1}$ la i-ésima observación de las variables independientes explicativas.

Sea Y_i la i-ésima observación de la variable respuesta.

Sea ε_i el i-ésimo error aleatorio.

Sea $i = 1,2,...,n$ el índice que corresponde al conjunto de tuplas o registros en la base de datos, que sirven cómo muestra de aprendizaje.

El modelo de regresión se puede escribir con notación matricial, así:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}_{nxp} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{px1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{nx1} \quad (2)$$

Es común la presentación del modelo en forma resumida:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3)$$

Donde $\mathbf{Y}$ es el vector de respuestas, $\mathbf{X}$ la matriz de constantes o variables independientes, $\boldsymbol{\beta}$ el vector de coeficientes y $\boldsymbol{\varepsilon}$ es el vector de error aleatorio.

Se dice que un modelo de regresión es lineal por la linealidad en los parámetros o coeficientes del modelo de regresión, y no por una restricción impuesta sobre las variables independientes que pueden explicar a Y. Por lo anterior, los paquetes estadísticos ofrecen varias alternativas de transformación de las variables independientes, como $X^2, \sqrt{X}$ ó e^X para ser aplicables antes de realizar un análisis de regresión.

La regresión lineal simple es una particularización de la regresión lineal multivariante donde se considera un solo regresor o variable independiente. Se parte de un conjunto de datos o pares ordenados que se pueden representar en el plano cartesiano, modelo gráfico conocido como Diagrama de Dispersión. Al observar la distribución de los puntos en el plano cartesiano, se puede intuir la existencia de una relación entre el conjunto de datos. Una relación puede ser representada por una línea recta, como se ilustra en la Figura 4.

En la Figura 4, se observa que los puntos de la gráfica no corresponden exactamente a la ecuación de la línea recta, al aceptarse que parte de la variación en $\hat{Y}$ no es explicada por la variable dependiente, sino por la naturaleza misma del fenómeno bajo estudio, por errores en las mediciones o por cualquier otra fuente de imperfección en los datos observados. Esta componente aleatoria se representa por ε_i en la ecuación (1).

La ecuación de la línea recta que ajusta los datos en la Figura 4 es:

$$\hat{Y} = \beta_0 + \beta_1 X \qquad (4)$$

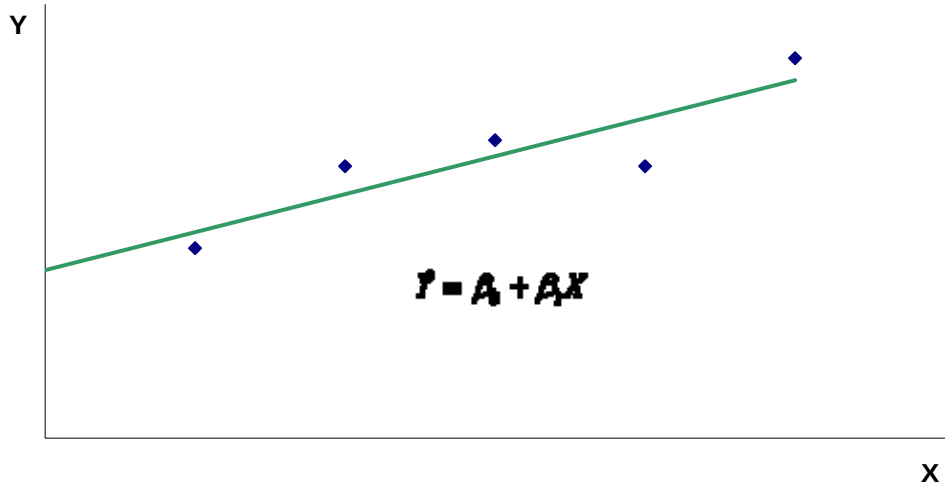


Figura 4. Diagrama de dispersión con línea de tendencia.

Los parámetros del modelo de regresión lineal simple son β_0, β_1 que corresponden al intercepto y a la pendiente de la línea de tendencia que ajusta los datos respectivamente. La representación geométrica del modelo de regresión con una variable independiente puede ser una línea recta o una curva cuando sobre la variable independiente se realiza una transformación. El modelo con dos variables independientes se puede representar por un plano o una superficie. No es fácil una representación geométrica del modelo de regresión lineal con más de dos variables independientes.

En el modelo de regresión lineal multivariante el parámetro β_0 es el intercepto (de la línea, plano, curva, hiperplano o hipercurva, dependiendo del número de variables independientes y de las posibles transformaciones); si el alcance del modelo incluye al intercepto, el parámetro β_0 es la media de la distribución de la variable de respuesta Y, cuando las demás variables explicativas son cero. Los coeficientes de regresión $\beta_1, \beta_2, ..., \beta_{p-1}$ miden el cambio esperado en la variable de respuesta Y, por unidad de cambio en la correspondiente variable explicativa cuando las demás se mantienen constantes. En el caso de no existir linealidad en las variables explicativas, la interpretación de los coeficientes de regresión puede ser mucho más compleja y dependerá de la forma final del modelo.

El problema central del análisis de regresión consiste en encontrar los estimadores más apropiados de los parámetros β_i, utilizando los datos observados. El científico alemán Gauss, propuso estimar los parámetros β_i minimizando la suma de los cuadrados de las desviaciones o las diferencias entre valores observados y ajustados. Este criterio para estimar los coeficientes de regresión se conoce como método de mínimos cuadrados (Montgomery y Runger, 2003). Sin embargo, también existen otros métodos de estimación de los parámetros como el método de máxima verosimilitud. Para un modelo de regresión donde los errores se distribuyen normalmente, estos dos métodos coinciden, por el teorema de Gauss-Markov, y los estimadores de los parámetros tienen las propiedades de ser estimadores insesgados, consistentes y suficientes (Weisberg, 2005).

La función de mínimos cuadrados es:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \left(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_{p-1} X_{i,p-1}\right)\right)^2 \qquad (5)$$

Al realizar las derivadas correspondientes e igualar a cero, resulta el sistema de ecuaciones a resolver:

$$\mathbf{X'X\hat{\beta} = X'Y} \qquad (6)$$

La expresión (6) se conoce como las ecuaciones normales en forma matricial de las cuales se obtiene, el estimador de mínimos cuadrados de $\hat{\beta}$ como (Montgomery y Runger, 2003):

$$\mathbf{\hat{\beta} = \left(X'X\right)^{-1} X'Y} \qquad (7)$$

3.1.1. Criterios de validación para la regresión lineal

Después de construir un modelo de regresión, se debe validar la bondad de ajuste con el fin de poderlo utilizar para describir o predecir valores no sólo futuros, sino para otros valores en las variables independientes no observados o medidos. Cuando hay suficientes datos, se utiliza un subconjunto de ellos, no considerados en el ajuste, para evaluar la capacidad predictiva del modelo. En otros casos, sólo es posible verificar la significación estadística de las variables explicativas, de manera global o particular, y el grado de cumplimiento de los supuestos impuestos al modelo de regresión lineal.

Una de las suposiciones del modelo de regresión lineal multivariante es la independencia de las variables explicativas, cuando esto no se cumple en su totalidad, se le suele denominar "multicolinealidad", ésta puede generar problemas en la estimación de los coeficientes de regresión y sobre la aplicabilidad general del modelo estimado. Aunque exista una relación estadística entre la variable respuesta y el conjunto de variables explicativas, muchos de los coeficientes de regresión estimados individualmente no son estadísticamente significativos (Neter *et al.*, 1996).

Significancia de la regresión lineal

La prueba para la significancia de la regresión determina si existe una relación lineal entre la variable de respuesta Y y un subconjunto de las variables de regresión. La hipótesis nula expresa que $\beta_1 = \beta_2 =,...,\beta_{p-1} = 0$ y rechazar dicha hipótesis implica que al menos un coeficiente de regresión sea diferente de cero, lo que a su vez significa, que al menos una de las variables de regresión tiene una contribución significativa en la variable respuesta Y (Neter *et al.*, 1996; Montgomery y Runger, 2003). El estadístico de prueba F_o se utiliza para determinar si la relación observada entre la variable respuesta y las variables explicativas se produce por azar. La hipótesis nula se debe rechazar si el valor calculado del estadístico de prueba es mayor que el comparado con el valor teórico de la distribución F al nivel de confianza establecido.

La prueba de significancia de la regresión lineal se realiza por medio de un análisis de varianza y se parte de las siguientes definiciones (Neter *et al.*, 1996; Montgomery y Runger, 2003).

Suma de los cuadrados de la regresión:

$$SS_R = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 \qquad (8)$$

Donde $\overline{Y}$ es el valor promedio de Y_i.

Suma de los cuadrados de los errores:

$$SS_E = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \quad (9)$$

Media de los cuadrados de la regresión:

$$MS_R = \frac{SS_R}{p-1} \quad (10)$$

Donde $p-1$ es el número de variables independientes de la regresión, también conocido como grados de libertad de la regresión.

Media de los cuadrados de los errores:

$$MS_E = \frac{SS_E}{n-p} \quad (11)$$

Donde $n-p$ son los grados de libertad del error o el residuo.

Estadístico de prueba:

$$F_o = \frac{MS_R}{MS_E} \quad (12)$$

La hipótesis nula $\beta_1 = \beta_2 =,..., \beta_{p-1} = 0$, se rechaza si el estadístico de prueba F_o es mayor que $f_{\alpha,p-1,n-p}$, de la distribución F, siendo α el nivel de confianza, $p-1$ los grados de libertad 1 y $n-p$ los grados de libertad 2.

Rechazar la hipótesis nula con el estadístico de prueba F, no implica de manera necesaria que la relación determinada por el modelo sea significativa, por esto, se requieren más pruebas, como la prueba sobre los coeficientes individuales de regresión. Estas pruebas son útiles para determinar el valor potencial de cada una de las variables de regresión del modelo (Montgomery y Runger, 2003).

Significancia de los coeficientes de regresión lineal

La hipótesis nula para la prueba de significancia del coeficiente de regresión j expresa $\beta_j = 0$. Si no se rechaza la hipótesis nula, la variable independiente correspondiente al coeficiente de regresión puede eliminarse del modelo. Esta prueba también se conoce como prueba parcial o marginal, y utiliza como estadístico de prueba a T_o.

El estadístico de prueba se define como la estimación del coeficiente de regresión j dividido por el error estándar del coeficiente de regresión.

El error estándar del coeficiente de regresión j, es:

$$s\{\hat{\beta}_j\} = \sqrt{MS_E \cdot C_{jj}} \quad (13)$$

Donde C_{jj} es el elemento de la diagonal de la matriz de varianzas y covarianzas $(\mathbf{X'X})^{-1}$ que corresponde a $\hat{\beta}_j$.

El estadístico de prueba para el coeficiente de regresión j es:

$$T_o = \frac{\hat{\beta}_j}{s\{\hat{\beta}_j\}} \quad (14)$$

La hipótesis nula $\beta_j = 0$, se rechaza cuando el valor absoluto del estadístico de prueba calculado es mayor que $t_{\alpha/2,n-p}$, de la distribución T, siendo $\alpha/2$ el nivel de confianza y $n-p$ los grados de libertad.

3.1.2. Coeficientes de regresión estandarizados

En un modelo de regresión lineal, la magnitud de los coeficientes de regresión esta determinada por las unidades de medida o escala de la variable independiente correspondiente. Por lo anterior, la magnitud del coeficiente de regresión no resalta la importancia relativa de cada variable independiente.

La determinación de coeficientes de regresión adimensionales puede facilitar la determinación de la importancia relativa de cada una de las variables independientes, estos coeficientes adimensionales se conocen como coeficientes de regresión estandarizados (Montgomery y Runger, 2003).

La relación existente entre los coeficientes del modelo original β_j y los coeficientes estandarizados $\beta_{std,j}$ es:

$$\hat{\beta}_{std,j} = \hat{\beta}_j \cdot \sqrt{\frac{S_{jj}}{S_{yy}}} \quad (15)$$

Siendo S_{yy} es la suma total de cuadrados corregida, definida como:

$$S_{yy} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \quad (16)$$

Análogamente se define a S_{jj}

$$S_{jj} = \sum_{i=1}^{n} \left(X_{ij} - \overline{X_j}\right)^2 \quad (17)$$

Donde $j = 1,2,..., p-1$

Los coeficientes de regresión estandarizados facilitan la comparación de los coeficientes, sin embargo, no debe emplearse la magnitud del coeficiente de regresión estandarizado como la única medida de importancia de las variables de regresión (Montgomery y Runger, 2003).

3.2. Regresión logística multivariante

La regresión logística es un tipo especial de regresión que se utiliza para predecir y explicar una variable categórica binaria, ejemplo: *Éxito* o *fracaso,* en lugar de una medida dependiente cuantitativa; por esto, la variable de salida o explicada se convierte en una variable binaria con valores de cero (0) o uno (1), para representar un fracaso o un éxito, pertenecer a una categoría o a su complemento. Esta técnica estadística también se considera como una técnica de análisis discriminante que permite explicar alguna característica o analizar problemas con una variable independiente o la combinación de varias variables independientes (Draper y Smith, 1966; Hair *et al.*, 1999; Lopera, 2002). En esta técnica el investigador está interesado en la predicción y explicación de las relaciones que influyen en la categoría en que un objeto está situado (Hair, *et al.*, 1999).

Sea Y la variable de respuesta dicotómica y explicada por una combinación lineal de variables X, se puede considerar que $Y=1$ si el evento ocurre, o $Y=0$ si no lo hace. En este modelo, el problema es encontrar la estimación de la probabilidad de ocurrencia de un evento, dado que ha ocurrido X. En otras palabras, la variable respuesta representa la probabilidad condicional de ocurrencia de un evento. La variable Y_i se puede considerar como una variable aleatoria tipo Bernoulli, denotada por:

$$\begin{aligned} &P(Y_i=1)=\pi_i \\ &P(Y_i=0)=1-\pi_i \\ &E\{Y_i\}=\pi_i \end{aligned} \qquad (18)$$

La aplicación directa del modelo de regresión lineal multivariante expresado en la ecuación (1) no es apropiado en el caso de una variable respuesta dicotómica, dado que las estimaciones que se hagan usando un modelo de regresión lineal puede generar valores diferentes a cero, o a uno. Por lo anterior, es necesaria una transformación antes de realizar el ajuste. El modelo matemático de regresión, se puede plantear de la siguiente manera.

$$Ln\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\beta_1 X_{i,1}+\beta_2 X_{i,2}+\ldots+\beta_{p-1}X_{i,p-1}=\mathbf{X}_i\boldsymbol{\beta} \qquad (19)$$

Donde:

$$\mathbf{X}_i = \begin{vmatrix} 1 & X_{i,1} & X_{i,2} & \cdots & X_{i,p-1} \end{vmatrix}_{1xp}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{px1}$$

La ecuación (19) define una superficie de respuesta altamente compleja pero que puede ser tratada como un modelo de regresión lineal multivariante, por ser intrínsecamente lineal. Una de las características interesantes de la regresión logística es la relación que guarda con un parámetro de cuantificación de riesgo, conocido en la literatura como "odds ratio". El odds asociado es el cociente entre la probabilidad de que ocurra un suceso, frente a la probabilidad de que no ocurra.

Por propiedades matemáticas, se cumple que:

$$\pi_i = \frac{e^{\mathbf{X}_i\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}} = \left[1+e^{-\mathbf{X}_i\boldsymbol{\beta}}\right]^{-1} \qquad (20)$$

El plano cartesiano muestra que la gráfica de la función logística expresada en la ecuación (20), con una sola variable independiente, tiene forma de S. La gráfica tiene asíntotas en cero (0) y uno (1), como se ilustra en la Figura 5.

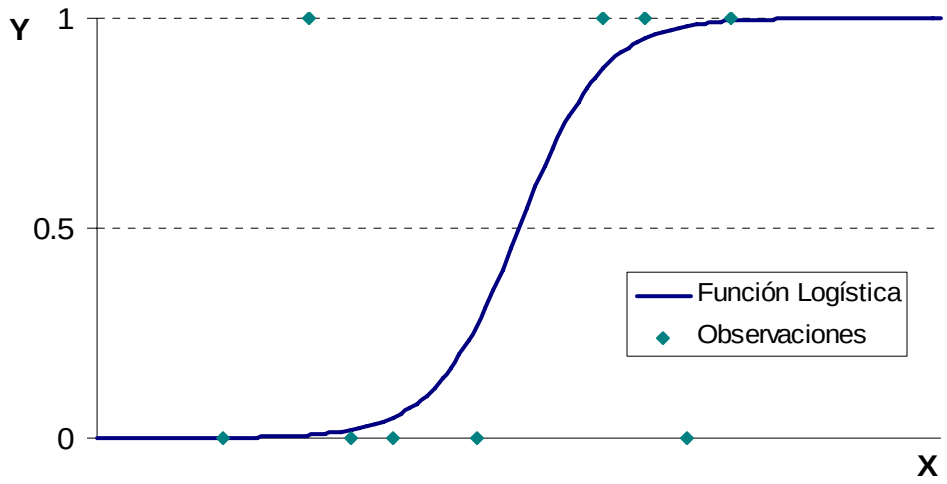


Figura 5. Gráfico de la función logística

El modelo de regresión logística se puede expresar como (Weisberg, 2005):

$$Y_i = E\{Y_i\} + \varepsilon_i$$
$$Y_i = \frac{e^{\mathbf{X}_i\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}} + \varepsilon_i \qquad (21)$$

3.2.1. Ajuste del modelo de regresión logística

El ajuste del modelo de regresión logística consiste en encontrar los estimadores más apropiados de los parámetros o coeficientes β_i, utilizando los datos observados. Existen varios procedimientos numéricos que permiten estimar los parámetros β_i, pero todos requieren de un uso intensivo de las capacidades de cálculo de los computadores. El método de mínimos cuadrados utilizado en la regresión lineal no se puede utilizar para estimar los parámetros β_i en regresión logística. En este caso, se requiere de procedimientos numéricos más complejos, un método usado con frecuencia es *El Algoritmo de Marquardt*, este algoritmo busca utilizar las mejores características del método de *Gauss-Newton* y el método de *steepest descent*, y ocupa un lugar intermedio entre estos dos métodos (Neter *et al.*, 1996).

A continuación se describen dos procedimientos iterativos de ajuste.

Método de Gauss-Newton
Este procedimiento es una búsqueda directa numérica y es aplicable para solucionar diversos problemas de regresión no lineal, utilizando métodos numéricos para buscar la solución de las ecuaciones de manera iterativa. El método de Gauss-Newton es también conocido como método de linealización usando expansiones en series de Taylor (Neter *et al.*, 1996).

Se reescribe la expresión (21):

$$Y_i = \pi(\mathbf{X}_i, \boldsymbol{\beta}) + \varepsilon_i \qquad (22)$$

El valor inicial para los parámetros del modelo $\beta_0, \beta_1, \beta_2, ..., \beta_{p-1}$ se puede obtener de estudios previos o de solucionar el modelo como si fuera una regresión lineal multivariante, denotándose como $b_0^{(0)}, b_1^{(0)}, b_2^{(0)}, ..., b_{p-1}^{(0)}$ (el número en el superíndice encerrado en paréntesis indicará la iteración).

De la expansión en series de Taylor, se obtiene:

$$\pi(\mathbf{X}_i, \boldsymbol{\beta}) \approx \pi(\mathbf{X}_i, \mathbf{b}^{(0)}) + \sum_{k=0}^{p-1} \left[\frac{\partial \pi(\mathbf{X}_i, \boldsymbol{\beta})}{\partial \beta_k} \right]_{\boldsymbol{\beta}=\mathbf{b}^{(0)}} \left(\beta_k - b_k^{(0)}\right) \qquad (23)$$

Para simplificar la notación:

$$\pi_i^{(0)} = \pi\left(\mathbf{X}_i, \mathbf{b}^{(0)}\right)$$
$$d_k^{(0)} = \left(\beta_k - b_k^{(0)}\right) \qquad (24)$$
$$D_{ik}^{(0)} = \left[\frac{\partial \pi\left(\mathbf{X}_i, \boldsymbol{\beta}\right)}{\partial \beta_k}\right]_{\boldsymbol{\beta}=\mathbf{b}^{(0)}}$$

$$\pi\left(\mathbf{X}_i, \boldsymbol{\beta}\right) \approx \pi_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} d_k^{(0)} \qquad (25)$$

Reorganizando la expresión (25), se obtiene:

$$\pi\left(\mathbf{X}_i, \boldsymbol{\beta}\right) - \pi_i^{(0)} \approx \sum_{k=0}^{p-1} D_{ik}^{(0)} d_k^{(0)} \qquad (26)$$

Se observa la semejanza con el modelo de regresión lineal multivariante por la linealidad de los términos $d_k^{(0)}$. La solución del sistema de ecuación se realiza por mínimos cuadrados y se mejora la estimación de los parámetros del modelo:

$$b_k^{(1)} = b_k^{(0)} + d_k^{(0)} \qquad (27)$$

Se calcula el criterio de mínimos cuadrados $SS_E{}^{(0)}$ y $SS_E{}^{(1)}$:

$$SS_E{}^{(j)} = \sum_{i=1}^{n} \left(Y_i - \pi\left(\mathbf{X}_i, \mathbf{b}^{(j)}\right)\right)^2 \qquad (28)$$

La convergencia del método se debe evaluar comparando las diferencias sucesivas:

$$SS_E{}^{(s+1)} - SS_E{}^{(s)}$$
$$\mathbf{b}^{(s+1)} - \mathbf{b}^{(s)} \qquad (29)$$

Cuando las diferencias de la expresión (29) son despreciables, el valor obtenido para $\mathbf{b}^{(s+1)}$ se convierte en la estimación de los parámetros del modelo de regresión logística $\boldsymbol{\beta}$.

Método de mínimos cuadrados iterativamente reponderados

La estimación de los parámetros $\boldsymbol{\beta}$, por máxima verosimilitud para el modelo de regresión logística se puede obtener por un procedimiento iterativo utilizando mínimos cuadrados ponderados (Neter *et al.*, 1996).

El valor inicial para los parámetros del modelo $\beta_0, \beta_1, \beta_2, ..., \beta_{p-1}$ se puede obtener de solucionar el modelo como si fuera una regresión lineal multivariante, denotándose como $b_0^{(0)}, b_1^{(0)}, b_2^{(0)}, ..., b_{p-1}^{(0)}$ (el número en el superíndice encerrado en paréntesis indicará la iteración).

Las probabilidades en la iteración cero se definen como:

$$\pi_i^{(0)} = \frac{e^{\mathbf{X}_i \mathbf{b}^{(0)}}}{1 + e^{\mathbf{X}_i \mathbf{b}^{(0)}}} \tag{30}$$

Los pesos en la iteración cero se definen como:

$$w_i^{(0)} = \pi_i^{(0)} \left(1 - \pi_i^{(0)}\right) \tag{31}$$

La nueva variable respuesta se define como:

$$Y_i^{(0)} = \mathbf{X}_i \mathbf{b}^{(0)} + \frac{Y_i - \pi_i^{(0)}}{w_i^{(0)}} \tag{32}$$

El siguiente paso es resolver el sistema de ecuaciones por mínimos cuadrados con la nueva variable respuesta como si se tratará de un análisis de regresión lineal multivariante. Los parámetros obtenidos pasan a ser la nueva estimación de los parámetros de la regresión lineal $b_0^{(1)}, b_1^{(1)}, b_2^{(1)}, ..., b_{p-1}^{(1)}$ y se repite el proceso hasta obtener convergencia con el mismo criterio del método anterior.

3.2.2. Criterios de validación para la regresión logística

Después de construir un modelo de regresión logística, se debe validar la bondad de ajuste con el fin de poderlo utilizar para describir o predecir valores no sólo futuros, sino para otros valores en las variables independientes no observados o medidos. Se verifica la significancia estadística de las variables explicativas, de manera global o particular, y el grado de cumplimiento de los supuestos impuestos al modelo de regresión logística.

Significancia de la regresión logística

Una de las pruebas para validar la bondad de ajuste de la regresión logística se basa en el "Deviance" del modelo. El "Deviance" del modelo ajustado compara el log de verosimilitud del modelo ajustado con el de un modelo saturado, y se define como:

$$DEV\left(\mathbf{X}_1, ... \mathbf{X}_{p-1}\right) = -2 \sum_{i=1}^{n} \left[Y_i \cdot Ln\left(\hat{\pi}_i\right) + \left(1 - Y_i\right) \cdot Ln\left(1 - \hat{\pi}_i\right) \right] \tag{33}$$

Sea $\hat{\pi}_i = \dfrac{e^{\mathbf{X}_i\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}$

Sea $\mathbf{X}_i\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_{p-1} X_{i,p-1}$

La hipótesis nula es $E\{Y\} = \left[1+e^{-\mathbf{X}\boldsymbol{\beta}}\right]^{-1}$ y se rechaza, sí el "Deviance" del modelo saturado $DEV\left(\mathbf{X}_1,...\mathbf{X}_{p-1}\right)$ es mayor que $\chi^2{}_{\alpha,n-p}$, de la distribución χ^2, siendo α el nivel de confianza y $n-p$ los grados de libertad. Rechazar la hipótesis nula significa que no existe suficiente evidencia estadística para asegurar que el modelo de regresión logística es adecuado, o posiblemente las variables consideradas no están explicando la variable respuesta.

Significancia de los coeficientes de regresión logística

La hipótesis nula para la prueba de significancia del coeficiente de regresión j expresa $\beta_j = 0$. Si no se rechaza la hipótesis nula, la variable independiente correspondiente al coeficiente de regresión puede eliminarse del modelo.

El cálculo del "Deviance" para el modelo reducido sin el coeficiente de regresión j es similar al cálculo del "Deviance" del modelo ajustado:

$$DEV\left(\mathbf{X}_1,...,\mathbf{X}_{j-1},\mathbf{X}_{j+1},...,\mathbf{X}_{p-1}\right) = -2\sum_{i=1}^{n}\left[Y_i \cdot Ln\left(\hat{\pi}_i{}^{R(j)}\right) + \left(1-Y_i\right)\cdot Ln\left(1-\hat{\pi}_i{}^{R(j)}\right)\right] \qquad (34)$$

Donde el superíndice $R(j)$ hace referencia al modelo reducido que no considera el coeficiente de regresión j, por lo tanto $\hat{\pi}_i{}^{R(j)}$ y $\mathbf{X}_i\boldsymbol{\beta}^{R(j)}$ se definen como:

$$\hat{\pi}_i{}^{R(j)} = \frac{e^{\mathbf{X}_i\boldsymbol{\beta}^{R(j)}}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}^{R(j)}}}$$

$$\mathbf{X}_i\boldsymbol{\beta}^{R(j)} = \beta_0 + \beta_1 X_{i,1} + ... + \beta_{j-1} X_{i,j-1} + \beta_{j+1} X_{i,j+1} + ... + \beta_{p-1} X_{i,p-1}$$

La diferencia entre el "Deviance" del modelo reducido menos el "Deviance" del modelo completo se denota así:

$$DEV\left(\mathbf{X}_j \mid \mathbf{X}_1,...,\mathbf{X}_{j-1},\mathbf{X}_{j+1},...,\mathbf{X}_{p-1}\right) = DEV\left(\mathbf{X}_1,...,\mathbf{X}_{j-1},\mathbf{X}_{j+1},...,\mathbf{X}_{p-1}\right) - DEV\left(\mathbf{X}_1,...\mathbf{X}_{p-1}\right) \qquad (35)$$

La hipótesis nula $\beta_j = 0$, se rechaza si la diferencia entre los "Deviances" $DEV\left(\mathbf{X}_j \mid \mathbf{X}_1,...,\mathbf{X}_{j-1},\mathbf{X}_{j+1},...,\mathbf{X}_{p-1}\right)$ es mayor que $\chi^2{}_{\alpha,1}$ de la distribución χ^2, siendo $1-\alpha$ el nivel de confianza con un grado de libertad.

4. MODELO PARA LA VISUALIZACIÓN DE RESULTADOS

En el presente capítulo, se propone un modelo para la representación e interpretación de los resultados del análisis de regresión. El modelo para la visualización de resultados busca facilitar el enfoque de Inteligencia del Negocio y el proceso de descubrimiento de conocimiento para personas no expertas, aprovechar las habilidades del cerebro humano y convertir el proceso de análisis en un proceso intuitivo. Esto contribuye a que la extracción de la información y los análisis se conviertan en una operación rutinaria y semiautomática, y a su vez, contribuye con la popularización e implementación de la utilización de la información almacenada para apoyar la toma de decisiones.

Es común que un rostro se recuerde más fácilmente que un nombre, lo que se reafirma en varias investigaciones realizadas para medir la capacidad de reconocimiento de imágenes del cerebro humano. En el libro de Buzan y Buzan (1996), se hace referencia a un experimento realizado por R. Haber, con 2560 diapositivas, presentando una imagen cada diez segundos. En este experimento, se hizo necesario varias sesiones durante varios días, para realizar posteriormente una prueba de reconocimiento con una precisión en el reconocimiento de imágenes por persona entre el 85 al 95%. R. Haber realizó un segundo y tercer experimento con resultados idénticos, afirmando que: “Estos experimentos con estímulos visuales apuntan a que el reconocimiento de imágenes es esencialmente perfecto. Los resultados habrían sido probablemente los mismos si en vez de 2.500 imágenes hubiéramos usado 25.000”. El libro de Buzan y Buzan (1996) afirma que otros investigadores como R.S. Nickerson consiguieron resultados aún más sorprendentes con una precisión promedio en el reconocimiento de imágenes del 98 al 99.9%. El proverbio: “Una imagen vale más que mil palabras”, se puede justificar por las habilidades corticales que ésta estimula: colores, formas, líneas, dimensiones, texturas y especialmente la imaginación.

Se considera que la representación gráfica de los resultados de un análisis de regresión es una herramienta poderosa para la descripción de las posibles relaciones estadísticas entre las variables explicativas con la variable respuesta, porque “las imágenes suelen ser más evocativas, precisas y directas que las palabras, cuando se trata de realizar una amplia gama de asociaciones” (Buzan y Buzan, 1996).

En general, las personas tienen la habilidad de detectar patrones o tendencias al observar una gráfica o una información tabular con una pequeña cantidad de datos, pero al incrementarse la información, tanto en cantidad como en atributos, resulta insuficiente dicho procedimiento y se hace necesario el

Descubrimiento de Conocimiento en Bases de Datos, que incluye las técnicas de regresión lineal y logística multivariante.

La última etapa en el Descubrimiento de Conocimiento en Bases de Datos es la *Interpretación y Presentación del Conocimiento*. De esta etapa, depende en gran medida la utilización del conocimiento ganado para la toma de decisiones en una organización, dado que, si el conocimiento ganado no es organizado y presentado de modo que el tomador de decisiones pueda entenderlo, es muy probable que dicho conocimiento no sea tenido en cuenta a la hora de tomar decisiones.

4.1. Características del modelo de visualización de resultados

El modelo para la visualización de resultados busca representar los resultados esenciales del modelo estimado, basado en la interpretación de regresión por medio de pruebas estadísticas de bondad de ajuste para la regresión y para los coeficientes de regresión.

A continuación se enuncian las características relevantes del modelo para la visualización de resultados propuesto:

- Cada variable se representa en forma de caja con bordes redondeados y en ella el nombre personalizado de la variable.

- La variable respuesta se ubica en el centro de la imagen con un color de fondo diferente a las demás variables, para resaltarla como el centro del análisis.

- Las variables explicativas se ubican equidistantes alrededor de la variable respuesta, similar a un gráfico tipo radar.

- Si las pruebas de bondad de ajuste del coeficiente de regresión de una variable, indica que una variable es significativa (con el nivel de confianza seleccionado), se dibuja una línea entre la variable explicativa y la variable respuesta, en caso de no ser significativa, no se dibuja la línea y el fondo de la caja de la variable se muestra en color blanco.

- El color de las líneas de unión de las variables explicativas con la variable respuesta representa el signo de la relación. El color rojo representa una relación negativa, que simboliza, que al aumentar el valor de la variable explicativa disminuye el valor de la variable respuesta. El color azul representa una relación positiva, que simboliza, que al aumentar el valor de la variable explicativa aumenta el valor de la variable respuesta.

- Se presentan los coeficientes de la regresión sobre la línea de unión de las variables. El coeficiente independiente se presenta en el centro inferior del gráfico.

- Las líneas de unión de las variables explicativas con la variable respuesta representa la relación estadística entre dichas variables. Se espera que todas las variables explicativas significativas, aporten en forma diferente al modelo de regresión estimado, el efecto relativo de cada variable se representa en el tipo y grosor de la línea. El efecto relativo de cada variable explicativa sobre la variable respuesta, se representa con la utilización de tres categorías: Fuerte, promedio y débil.

    a) El efecto relativo débil se representa con una línea punteada y delgada.
    b) El efecto relativo promedio se representa con una línea continua y delgada.
    c) El efecto relativo fuerte se representa con una línea continua y gruesa.

Se optó por representar en forma vaga el efecto relativo por ser habitual en la comunicación humana la utilización de adjetivos vagos como débil o fuerte. Adicionalmente, el significado de cada una de las categorías elegidas se puede entender intuitivamente en el gráfico propuesto.

En la regresión lineal multivariante se utilizan los coeficientes de regresión estandarizados, dado que por comparación de magnitud se puede determinar la importancia relativa de cada variable explicativa, por lo tanto, se define el siguiente índice:

$$I_{Lin,j} = \frac{\left|\hat{\beta}_{std,j}\right|}{\sum_{i=1}^{p-1} \hat{\beta}_{std,i}} \qquad (36)$$

Sí el índice $I_{Lin,j} < 0.25$, la variable explicativa j ejerce un efecto relativo débil sobre la variable respuesta, en el caso, que el índice $I_{Lin,j} \geq 0.25$ y $I_{Lin,j} \leq 0.5$, la variable explicativa j ejerce un efecto relativo promedio sobre la variable respuesta. Finalmente, sí el índice $I_{Lin,j} > 0.5$, la variable explicativa j ejerce un efecto relativo fuerte sobre la variable respuesta.

En la regresión logística multivariante se utiliza el “Deviance” para el modelo reducido sin el coeficiente de regresión j, por lo tanto, se define el siguiente índice:

$$I_{Log,j} = \frac{\left|DEV\left(\mathbf{X}_1,...,\mathbf{X}_{j-1},\mathbf{X}_{j+1},...,\mathbf{X}_{p-1}\right)\right|}{\sum_{i=1}^{p-1}\left|DEV\left(\mathbf{X}_1,...,\mathbf{X}_{i-1},\mathbf{X}_{i+1},...,\mathbf{X}_{p-1}\right)\right|} \quad (37)$$

Sí el índice $I_{Log,j} > 0.5$, la variable explicativa j ejerce un efecto relativo débil sobre la variable respuesta, en el caso, que el índice $I_{Log,j} \geq 0.25$ y $I_{Log,j} \leq 0.5$, la variable explicativa j ejerce un efecto relativo promedio sobre la variable respuesta. Finalmente, sí el índice $I_{Log,j} < 0.25$, la variable explicativa j ejerce un efecto relativo fuerte sobre la variable respuesta.

Los coeficientes de regresión se presentan en el modelo para hacerlo más completo y que el usuario pueda inferir fácilmente el modelo matemático correspondiente al análisis, pero el modelo matemático es un resultado indispensable.

4.2. Ejemplificación del modelo de visualización de resultados

El modelo para la visualización de resultados se presenta por medio de un ejemplo, para esto se utilizará la base de datos “benchmark”, que contiene información sobre 398 automóviles (ITL, 2006). En los ejemplos del presente capítulo, no se considerarán seis tuplas o registros, donde el valor de la *Potencia* es nulo (no definido).

En la Figura 6, se muestra el modelo para la visualización de resultados del análisis de regresión. Es fácil reconocer la hipótesis que se esperaba verificar, donde *Rendimiento* es la variable dependiente o explicada, los atributos o variables independientes son la *Potencia*, el *Peso* y la *Aceleración*. El modelo para la visualización de resultados permite reconocer en forma intuitiva la regresión planteada.

La capacidad del cerebro humano para crear asociaciones a partir de una idea o concepto central con la precisión en el reconocimiento de imágenes, hace del modelo para la visualización de resultados una excelente alternativa. Por su riqueza expresiva, el modelo para la visualización de resultados, se puede considerar como un paso para facilitar que una persona sin necesidad de mayores conocimientos técnicos pueda comprender intuitivamente los resultados del análisis. El modelo para la visualización de resultados presentado en la Figura 6, permite concluir fácilmente que:

- No existe suficiente evidencia estadística para afirmar que la *Aceleración* incide en el *Rendimiento.*

- Existe una relación negativa y estadísticamente significativa entre la *Potencia* y el *Rendimiento*.

- Existe una relación negativa y estadísticamente significativa entre el *Peso* y el *Rendimiento*.

- Entre el efecto de la *Potencia* y el efecto del *Peso* sobre el *Rendimiento estimado*, el *Peso* es la variable más importante, en otras palabras, el cambio del *Peso* tiene un efecto mayor en el *Rendimiento estimado*.

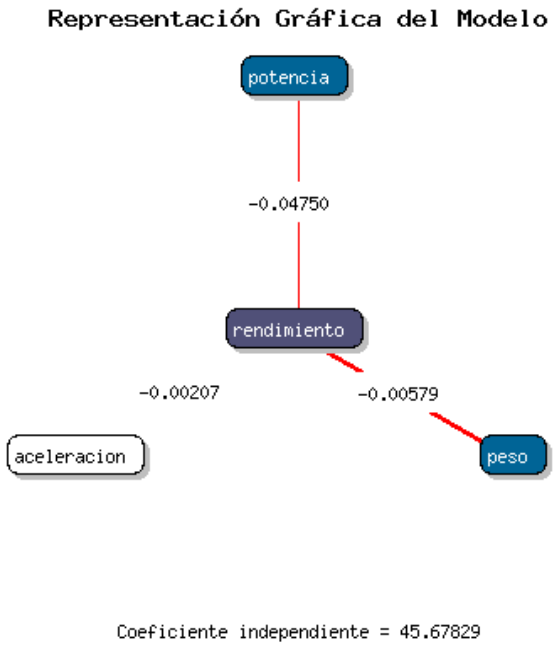


Figura 6. Modelo de visualización para describir la variable *Rendimiento*

El modelo matemático correspondiente al gráfico presentado en la Figura 6, con un nivel de confianza del 95% es:

$$RENDIMIENTO_{ESTIMADO} = (45.67829) + (-0.04750) \cdot POTENCIA + (-0.00579) \cdot PESO + (-0.00207) \cdot ACELERACIÓN$$

En este ejemplo, se observa que todas las variables explicativas presentan una relación negativa con el *Rendimiento Estimado*, dado que a mayor *Potencia* menor *Rendimiento*, a mayor *Peso* menor *Rendimiento*, a mayor *Aceleración* menor *Rendimiento*.

La Figura 6, muestra que la variable *Aceleración* no es una variable significativa en el modelo, por esto, se repite el análisis de regresión sin la

variable *Aceleración*, ahora el modelo final es más simple. El modelo matemático correspondiente, con un nivel de confianza del 95% es:

$$RENDIMIENTO_{ESTIMADO} = (45.64021) + (-0.04730) \cdot POTENCIA + (-0.00579) \cdot PESO$$

El alcance del modelo para el *Rendimiento Estimado* no incluye al intercepto, porque no tiene sentido, ni es posible, construir un automóvil con *Potencia* cero (0) y/o *Peso* cero (0), lo cual, implica que el coeficiente independiente no tiene significado práctico. El coeficiente de la variable explicativa *Potencia* dice que en promedio, por un caballo de fuerza que se le incremente a un automóvil, su *Rendimiento* es disminuido en 0.04730 millas por galón, manteniendo el *Peso* constante. Análogamente, el coeficiente de la variable explicativa *Peso* expresa que en promedio, por cada unidad de peso que se le incremente a un automóvil, su *Rendimiento* es disminuido en 0.00579 millas por galón, manteniendo la *Potencia* constante.

Se puede lograr un impacto mayor en el Descubrimiento de Conocimiento en Bases de Datos, si se presentan los resultados adecuados al usuario adecuado. Por ejemplo: Un gerente puede estar interesado sólo en conocer las relaciones existentes; en ese caso, el modelo de visualización es suficiente, otro sería el caso, si es un experto analista de datos.

La presentación de resultados con el modelo de visualización no implica dejar de presentar los resultados numéricos del ajuste del modelo, las gráficas para el análisis de los residuales o la matriz de varianzas y covarianzas. Se deben posibilitar todas las opciones porque el objetivo del análisis no siempre es el mismo.

Es importante resaltar que la interpretación de un análisis de regresión cambia cuando se cambia la variable respuesta, así se utilice el mismo conjunto de datos. En la Figura 7, se muestra el modelo para la visualización de resultados del análisis de regresión con el mismo conjunto de datos utilizados en el ejemplo anterior, pero la variable respuesta en este caso es la *Potencia*. En la Figura 6, la *Potencia* era la variable independiente, todo cambia cuando lo que se desea explicar es la *Potencia*. El modelo de visualización presentado en la Figura 7, permite concluir que:

- Existe una relación negativa y estadísticamente significativa entre el *Rendimiento* y la *Potencia*.
- Existe una relación positiva y estadísticamente significativa entre el *Peso* y la *Potencia*.
- Existe una relación negativa y estadísticamente significativa entre la *Aceleración* y la *Potencia*.

- El *Peso* es la variable más importante para explicar la variación de la *Potencia*.

- El efecto del *Rendimiento* es débil con respecto al *Peso* y la *Aceleración*.

- El efecto de la *Aceleración* es promedio con respecto al *Rendimiento* y el *Peso*.

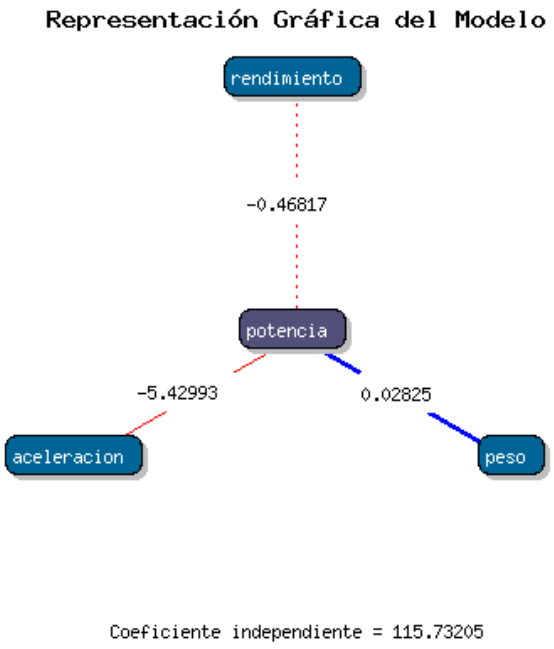


Figura 7. Modelo de visualización para describir la variable *Potencia*

El modelo matemático correspondiente al gráfico presentado en la Figura 7, con un nivel de confianza del 95% es:

$$POTENCIA_{ESTIMADA} = (115.73205) + (-0.46817) \cdot RENDIMIENTO + (0.02825) \cdot PESO + (-5.42993) \cdot ACELERACIÓN$$

En este ejemplo, Se observa que todas las variables explicativas son significativas; igual que en el ejemplo anterior, el alcance del modelo para la *Potencia Estimada* no incluye al intercepto. El coeficiente de la variable explicativa *Rendimiento* dice que en promedio, por cada milla por galón que se le incremente al rendimiento de un automóvil, su *Potencia* debe ser disminuida en 0.46817 caballos de fuerza, manteniendo *Peso* y *Aceleración* constantes. Análogamente, el coeficiente de la variable explicativa *Peso* dice que en promedio, por cada unidades de peso que se le incremente a un automóvil, su *Potencia* es aumentada en 0.02825 caballos de fuerza, manteniendo *Rendimiento* y *Aceleración* constantes, igualmente el coeficiente de la variable explicativa *Aceleración* dice que en promedio, por cada unidad en la medida de la aceleración máxima que se le incremente a un automóvil, su *Potencia* es

disminuida en 5.42993 caballos de fuerza, manteniendo *Peso* y *Rendimiento* constantes.

El modelo para la visualización de resultados puede ser utilizado con la misma validez para presentar e interpretar los resultados de la regresión lineal ó logística.

5. MODELADO DE LAS TÉCNICAS DE REGRESIÓN

La presente Tesis de Maestría busca la generación de herramientas que faciliten y promuevan el Descubrimiento de Conocimiento en Bases de Datos. Las técnicas de análisis de regresión lineal y logística multivariante, son consideradas como técnicas de Minería de Datos; y por sus características presentan ventajas frente a otras técnicas de Minería de Datos, por esto, el objetivo general de la Tesis de Maestría, es la incorporación de las técnicas de regresión lineal y logística multivariante en un gestor de bases de datos, en este caso un gestor de bases de datos de distribución libre.

A continuación se enuncian las características a resaltar de la regresión lineal y logística multivariante:

- Exploran los datos en forma supervisada en búsqueda de relaciones entre sus atributos. Las técnicas se utilizan con frecuencia y éxito en múltiples situaciones.

- Revela relaciones entre atributos y cuantifica dichas relaciones.

- La interpretación es simple y directa.

- Los supuestos son claros y fáciles de cumplir, adicionalmente el supuesto de normalidad se puede relajar a distribuciones simétricas, y se pueden lograr resultados de gran calidad con muestras de tamaños relativamente pequeñas.

- Los tiempos de cómputo utilizados en estas técnicas son relativamente pequeños, en comparación con otras técnicas de regresión.

- El modelo resultante es descriptivo y predictivo.

En resumen, las técnicas de regresión lineal y logística multivariante son técnicas válidas en diferentes enfoques de Minería de Datos, dado que se aplican a problemas de clasificación, de análisis de dependencias y problemas de predicción. Se puede decir que son técnicas necesarias más no suficientes en un proyecto de Descubrimiento de Conocimiento. El conjunto de características de estas técnicas son suficientes para justificar su utilización en la presente Tesis.

En la actualidad, los gestores de bases de datos son las principales herramientas para almacenar grandes cantidades de información, de manera organizada y estructurada; por lo anterior, es acertado pensar que los gestores de bases de datos deberían incorporar suficientes capacidades para analizar los

datos almacenados. La incorporación de capacidades de análisis en los gestores de bases de datos contribuye a mejorar la eficiencia total del proceso de Descubrimiento de Conocimiento en Bases de Datos al eliminar la necesidad de importar o vincular datos a otras herramientas de análisis como paquetes estadísticos.

La presente Tesis de Maestría se enmarca en el desafío de facilitar el proceso de Descubrimiento de Conocimiento en Bases de Datos, para lo cual se propone:

- Crear una aplicación Web inteligente para el planteamiento del análisis, interpretación y visualización de resultados.

- Incorporar dos técnicas de análisis multivariante en un gestor de bases de datos de distribución libre.

- Presentar los resultados con un modelo de visualización de los resultados de un análisis, y también mostrar los resultados en forma tradicional.

Lo anterior, busca la popularización del enfoque de Inteligencia del Negocio al simplificar el proceso de análisis, sin la creación de Bodegas de Datos o modelos multidimensionales, eliminado la necesidad de especialistas en el manejo de aplicaciones sofisticadas.

5.1. Incorporación de las técnicas de análisis multivariante

La propuesta de incorporación de las técnicas de análisis multivariante es independiente del sistema gestor de bases de datos, siempre y cuando el lenguaje subyacente del sistema gestor permita la creación de funciones o procedimientos almacenados. La propuesta requiere la creación de cinco funciones, la adición de tres tablas y la creación de dos tipos de datos:

- La función para la regresión lineal
- La función para la regresión logística
- La función para encontrar los valores de la distribución F
- La función para encontrar los valores de la distribución T
- La función para encontrar los valores de la distribución χ^2
- La tabla estadística con la distribución F
- La tabla estadística con la distribución T
- La tabla estadística con la distribución χ^2
- El tipo de dato con todos los resultados de la regresión lineal
- El tipo de dato con todos los resultados de la regresión logística

La Figura 8, presenta el diagrama de flujo de la función de regresión lineal, esta función tiene cinco argumentos de entrada: Un vector con todos los datos para el análisis, un valor entero con el número de coeficientes de regresión, un valor entero con el número total de observaciones, un valor lógico – falso o verdadero – que indica la inclusión del coeficiente independiente y, por último, un valor real con el nivel de significancia para la validación del modelo y de los coeficientes de regresión.

En el capítulo 3, se describieron dos métodos para ajustar los parámetros de la regresión logística y se mencionaron que existen otros métodos o combinaciones de ellos. Los dos métodos descriptos son válidos, pero era necesario seleccionar uno. La selección del método a utilizar en el prototipo desarrollado se realizó por comparación de los resultados por medio de dos ejemplos.

Se consideran dos factores para seleccionar el método, el primero es la sumatoria del valor absoluto de los residuos y el segundo es la sumatoria de los cuadrados de los residuos.

$$S_E = \sum_{i=1}^{n} |\varepsilon_i| \tag{38}$$

$$SS_E = \sum_{i=1}^{n} (\varepsilon_i)^2 \tag{39}$$

Se comparan los resultados de dos conjuntos de datos, los datos están disponibles en el CD-ROM anexo al libro de Neter *et al.* (1996). El primer conjunto de datos tomado del problema 6 del capítulo 14, conjunto de 30 observaciones para una regresión logística simple. El segundo conjunto de datos tomado de la tabla 3 del capítulo 14, conjunto de 98 observaciones para una regresión logística multivariante. En la Tabla 2, se presenta un resumen de los resultados obtenidos.

Tabla 2. Métodos de ajuste para la regresión logística

Primer Conjunto de Datos - Regresión Logística Simple		
Procedimiento de Solución	S_E	SS_E
Paquete Estadístico R	13.01334	6.50147
Método de Gauss-Newton	12.97950	6.50077
Método de Mínimos Cuadrados Ponderados	13.03987	6.50309
Segundo Conjunto de Datos - Regresión Logística Multivariante		
Procedimiento de Solución	S_E	SS_E
Solución del libro de Neter *et al.* (1996)	33.73231	17.03173
Método de Gauss-Newton	34.46464	16.98114
Método de Mínimos Cuadrados Ponderados	33.00477	17.21741

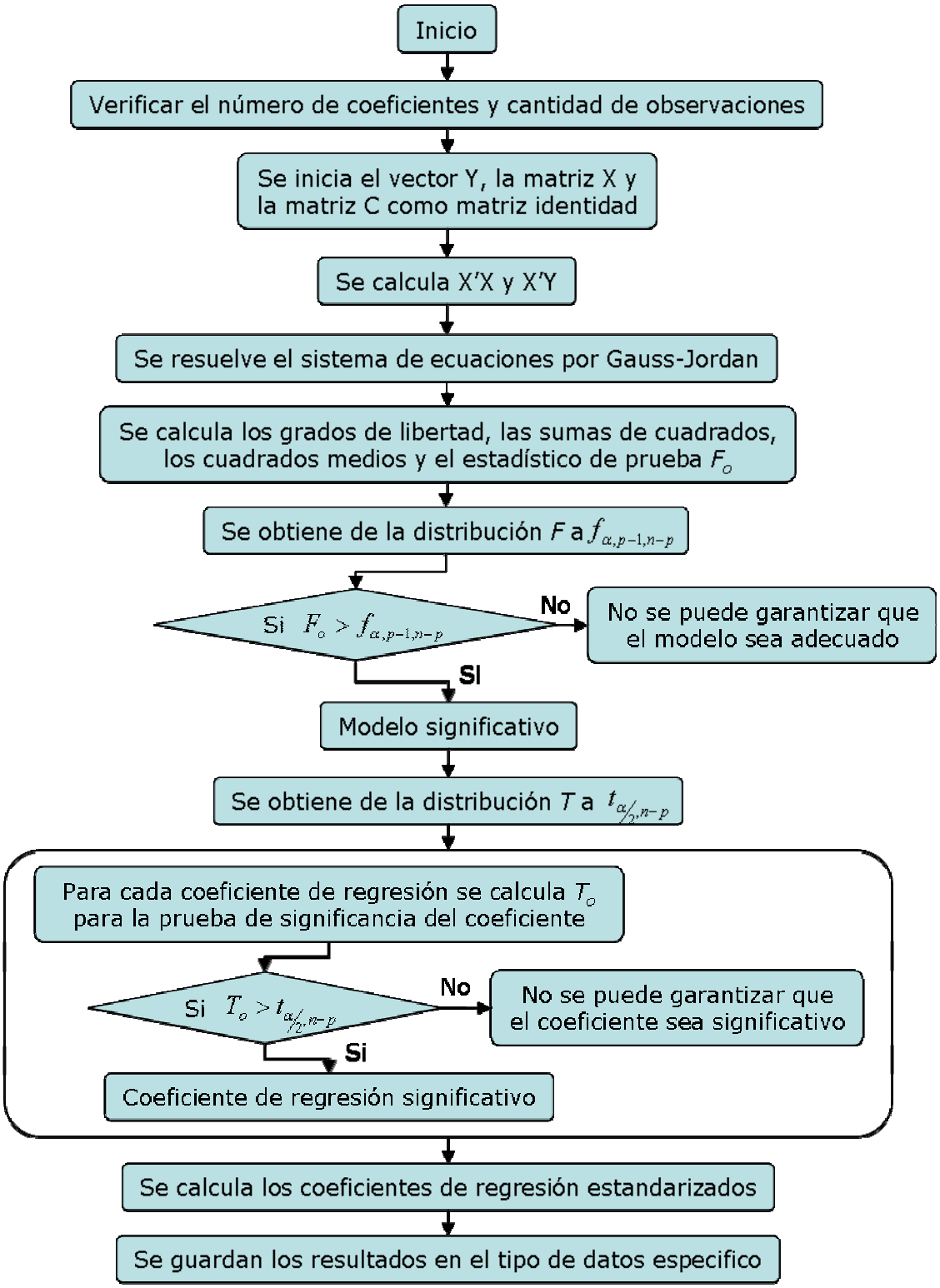


Figura 8. Diagrama de flujo de la función de regresión lineal

Todos los procedimientos de solución mencionados en la Tabla 2 son válidos. Al realizar una comparación entre los valores reportados se observa que en el primer conjunto de datos, el método de Gauss-Newton fue el procedimiento de ajuste que mejor optimizó los dos factores de selección definidos en las expresiones (38) y (39). En el segundo conjunto de datos, el método de Gauss-Newton fue el procedimiento de ajuste que mejor optimizó la sumatoria de los cuadrados del residuo, mientras que, el método de mínimos cuadrados ponderados fue el procedimiento de ajuste que mejor optimizó la sumatoria del valor absoluto de los residuos. La solución planteada en el ejemplo del libro de Neter *et al.* (1996) se encuentra en un punto intermedio para ambos factores de selección pero en ambos casos más cercana a la solución del método de Gauss-Newton. De la comparación de los valores de la Tabla 2, se selecciona el método de Gauss-Newton como el procedimiento de ajuste para utilizar en el prototipo desarrollado.

La Figura 9, presenta el diagrama de flujo de la función de regresión logística basada en el método de Gauss-Newton, esta función tiene cinco argumentos de entrada: Un vector con todos los datos para el análisis, un valor entero con el número de coeficientes de regresión, un valor entero con el número total de observaciones, un valor lógico – falso o verdadero – que indica la inclusión del coeficiente independiente y, por último, un valor real con el nivel de significancia para la validación del modelo y de los coeficientes de regresión.

Las tablas estadísticas utilizadas en las pruebas de significancia planteadas en las técnicas de regresión lineal y logística, se describen con la utilización de un modelo lógico de datos. Un modelo lógico de datos en este caso describe la estructura de las tablas sin ligarlas a una plataforma específica (Jiménez, 1999). Todos los atributos o campos definidos en las tablas estadísticas tienen la característica de no permitir valores nulos o también llamados valores no definidos. El término G.L. significa grados de libertad. En la Tabla 3, se presenta el modelo lógico de datos de la tabla estadística de la distribución F. En la Tabla 4, se presenta el modelo lógico de datos de la tabla estadística de la distribución T. Por último, en la Tabla 5, se presenta el modelo lógico de datos de la tabla estadística de la distribución χ^2. Los datos ingresados en las tablas se obtienen del Apéndice A del libro de Montgomery y Runger (2003) o de libros de estadística que presenten las tablas estadísticas utilizadas.

Se requiere una función que devuelva el valor $f_{\alpha,\nu1,\nu2}$, de la distribución F, esta función tendrá tres argumentos de entrada: Alfa (α), los grados de libertad del numerador ($\nu1$) y los grados de libertad del denominador ($\nu2$). La función debe buscar o interpolar el valor $f_{\alpha,\nu1,\nu2}$ de los datos de la tabla estadística de la distribución F.

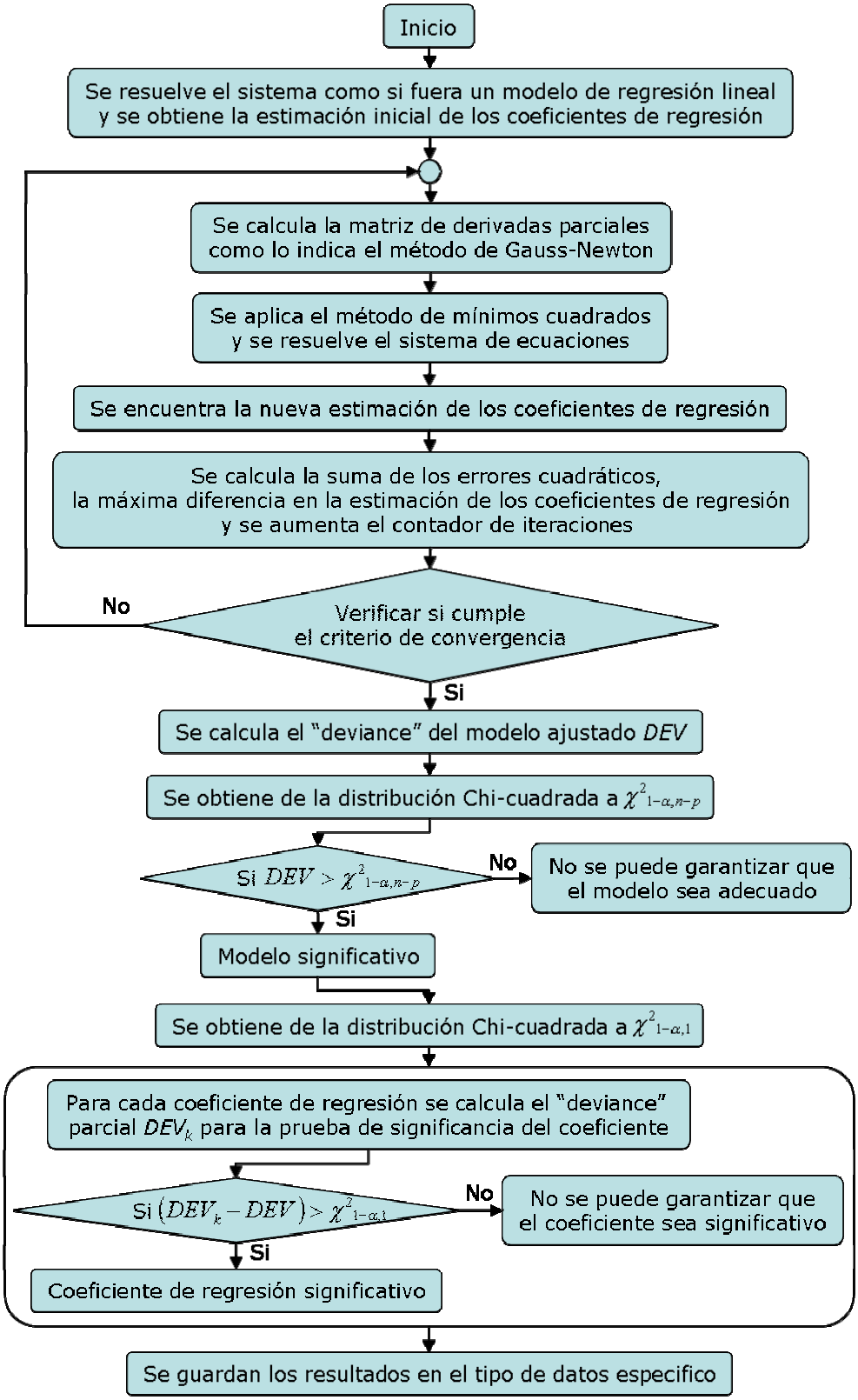


Figura 9. Diagrama de flujo de la función de regresión logística

Se requiere una función que devuelva el valor $t_{\alpha,\nu}$ de la distribución T, esta función tendrá dos argumentos de entrada: Alfa (α) y los grados de libertad de la distribución (ν). La función debe buscar o interpolar el valor $t_{\alpha,\nu}$ de los datos de la tabla estadística de la distribución T.

Se requiere una función que devuelva el valor $\chi^2{}_{\alpha,\nu}$ de la distribución χ^2, esta función tendrá dos argumentos de entrada: Alfa (α) y los grados de libertad de la distribución (ν). La función debe buscar o interpolar el valor $\chi^2{}_{\alpha,\nu}$ de los datos de la tabla estadística de la distribución χ^2.

Tabla 3. Modelo lógico de la tabla estadística de la distribución F

Atributo	Descripción	Clave	¿Valor único?	Tipo	Ejemplo
alfa	α	Primaria	Si	Real	0.05
k1	G.L. del numerador			Entero	15
k2	G.L. del denominador			Entero	28
f	Distribución F		No	Real	2.04

Tabla 4. Modelo lógico de la tabla estadística de la distribución T

Atributo	Descripción	Clave	¿Valor único?	Tipo	Ejemplo
alfa	α	Primaria	Si	Real	0.025
gl	G.L.			Entero	15
t	Distribución T		No	Real	2.131

Tabla 5. Modelo lógico de la tabla estadística de la distribución χ^2

Atributo	Descripción	Clave	¿Valor único?	Tipo	Ejemplo
alfa	α	Primaria	Si	Real	0.05
gl	G.L.			Entero	21
x2	Distribución χ^2		No	Real	32.67

La incorporación de las técnicas de análisis multivariante en un gestor de bases de datos no es suficiente en los objetivos propuestos en el capítulo 1 de la presente Tesis de Maestría, adicional, se debe plantear los procedimientos necesarios para la construcción de una interfaz inteligente que facilite la selección de los datos y la presentación e interpretación de los resultados del análisis. En la siguiente sección se presenta el proceso general para el descubrimiento de conocimiento con las técnicas multivariantes incorporadas.

5.2. Modelo de comportamiento del sistema

El modelo de comportamiento del sistema con un alto nivel de abstracción describe a grandes rasgos el proceso general para descubrir conocimiento con la utilización de las técnicas de análisis multivariantes incorporadas en un sistema gestor de bases de datos.

La Figura 10, presenta el modelo de comportamiento del sistema o proceso general para descubrir conocimiento con la utilización de las técnicas de análisis multivariantes incorporadas.

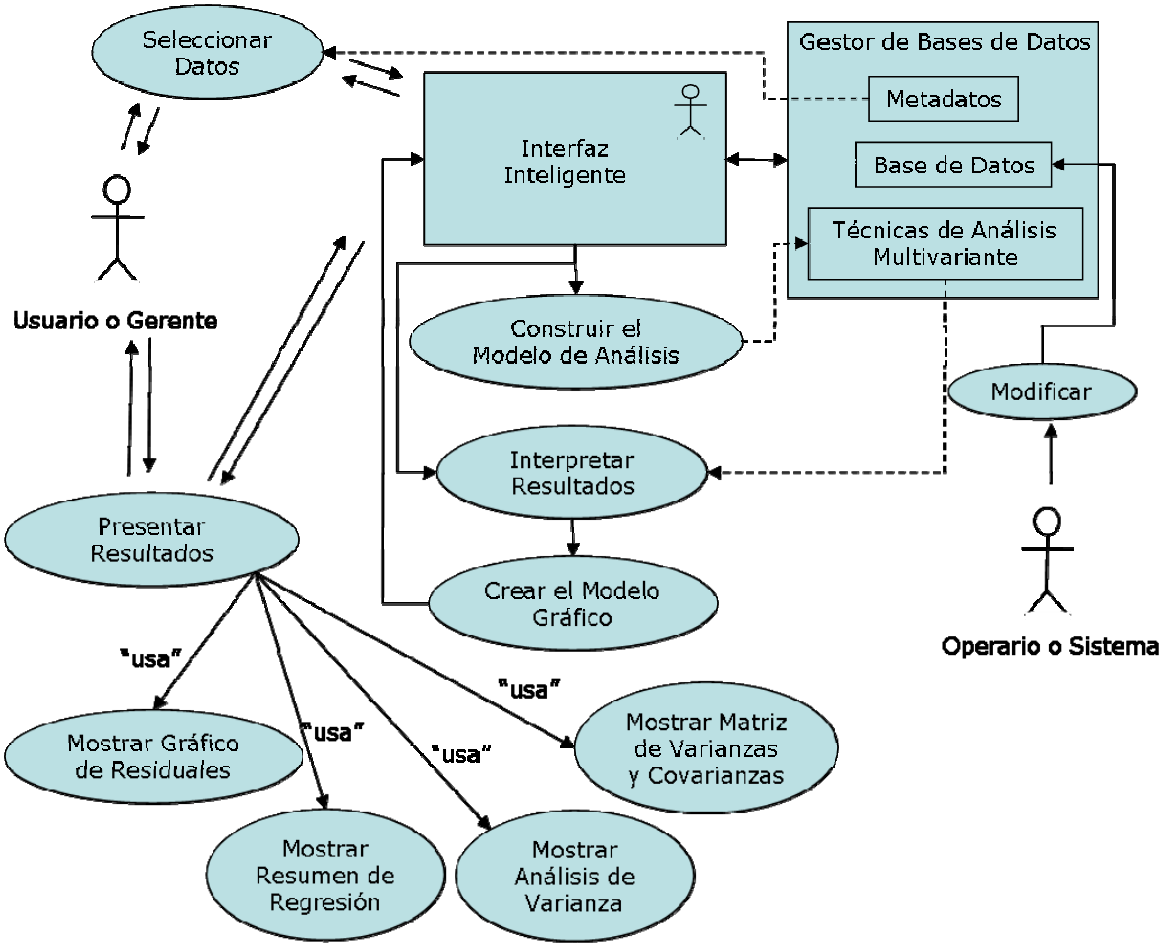


Figura 10. Modelo de comportamiento del sistema

A continuación se explica brevemente el modelo de comportamiento del sistema:

- El usuario o gerente selecciona los atributos para el análisis con ayuda de la interfaz inteligente y los metadatos o diccionario de datos, el usuario puede condicionar o filtrar los datos y establecer las demás

opciones para el análisis, siempre hay una doble vía de comunicación entre el usuario y la interfaz buscando facilitar el proceso al usuario.

- Se realiza el análisis léxico, sintáctico y semántico. Se construye el modelo de análisis con la utilización del lenguaje SQL. Se verifica que la petición esté correctamente formulada.

- Se solicita el análisis al gestor de bases de datos por medio de las funciones y procedimientos almacenados.

- Los resultados del gestor de bases de datos son utilizados por la interfaz inteligente para interpretar los resultados.

- Se crea el modelo de visualización que condensa los resultados para facilitar la asimilación del nuevo conocimiento.

- Se crea el gráfico de residuales que corresponde a la diferencia entre la variable respuesta observada menos la variable respuesta estimada.

- Se crean las tablas correspondientes al resumen de análisis de regresión y prueba de significancia de la regresión.

- Se presentan los resultados con la ayuda de una interfaz amigable que le permita al directivo tomar decisiones acertadamente.

- La base de datos puede ser modificada por cualquier usuario válido o por el sistema.

5.3. Propiedades de la solución planteada

A continuación se listan las propiedades deseables y obtenidas en la solución planteada.

- *Generalidad*: La propuesta aquí planteada es independiente del sistema gestor de bases de datos, el requisito solicitado es que el lenguaje subyacente del sistema gestor permita la creación de funciones o procedimientos almacenados.

- *Correctitud*: Hace referencia a que los resultados sean los esperados. En el capítulo 6, se muestran ejemplos de aplicación y los resultados han sido comparados con diferentes herramientas, entre las cuales, se encuentra el paquete estadístico R; de la comparación se obtiene que los resultados de la regresión lineal multivariante es exacta, mientras que la regresión logística presenta pequeñas divergencias que son atribuidas al

procedimiento numérico de solución utilizado. Se programaron dos procedimientos de solución para la regresión logística y finalmente se optó por el método de Gauss-Newton al obtener mejores resultados.

- *Exhaustivo*: Hace referencia a que se proporcione todos los resultados esperados por los usuarios interesados. Se presentan todos los resultados que se consideran necesarios para diferentes tipos de usuarios.

- *Robustez*: Hace referencia al buen manejo de las excepciones para evitar fallas en los procesos. Se realizaron un conjunto de pruebas para observar el comportamiento del prototipo desarrollado siempre con los resultados esperados.

- *Buen ajuste*: Hace referencia a que se realicen las pruebas de bondad de ajuste necesarias y se presenten los resultados de dichas pruebas. Se realizan pruebas de bondad de ajuste para la regresión y para cada uno de los coeficientes de regresión, se interpretan los resultados y se presentan las recomendaciones correspondientes.

- *Eficiencia*: Hace referencia a que el modelo debe ser eficiente en sus cálculos computacionales y en el manejo de la memoria. Se realizó un seguimiento de los tiempos de cómputo en diversos ejemplos y se obtuvo que los análisis de regresión son muy rápidos, en general, menos de un segundo para obtener la respuesta del sistema gestor de bases de datos. En general se puede considerar que el sistema es eficiente.

- *Simplicidad*: Hace referencia a que el modelo debe ser simple y comprensible. Las técnicas de análisis de regresión utilizadas tienen las características de ser simples y con interpretación clara. El modelo propuesto conserva dichas características.

- *Amigabilidad*: Hace referencia a varios aspectos relacionados con facilitar su uso a cualquier tipo de usuario. El prototipo desarrollado es amigable en las interfaces utilizadas y adicionalmente, interpreta los resultados y los presenta con diferentes formatos: Gráfico y tabular.

6. PROTOTIPO DESARROLLADO

PostgreSQL es el gestor de bases de datos de código abierto en la categoría de las bases de batos identificadas como objeto-relacionales. PostgreSQL se inició en la Universidad de California en Berkeley y fue pionera en muchos de los conceptos de bases de datos relacionales orientadas a objetos que ahora empiezan a estar disponibles en algunas bases de datos comerciales (Lockhart, 1999). PostgreSQL es un descendiente de dominio público y código abierto del código original de Berkeley con grandes proyecciones de desarrollo. Por lo anterior, PostgreSQL fue seleccionado como gestor de bases de datos de distribución libre, por el equipo de trabajo del mega-proyecto, que busca la generación de un conjunto de herramientas gerenciales para facilitar la aplicación de un enfoque de Inteligencia de Negocios y promover el Descubrimiento de Conocimiento en Bases de Datos. La presente Tesis de Maestría se enmarca en el mega-proyecto, por lo cual, se decidió utilizar PostgreSQL para el desarrollo del prototipo.

En PostgreSQL se crearon y poblaron las siguientes tablas con la información de la distribución F, la distribución t y la distribución ji-cuadrada, también se crearon las respectivas funciones para buscar o interpolar un valor en las distribuciones estadísticas. Se crearon las funciones de regresión lineal y logística multivariante, con las respectivas estructuras o tipos de datos definidos por el usuario para establecer la información de salida de cada una de las funciones.

La aplicación Web se desarrolla con la utilización de HTML, PHP y JAVASCRIPT como lenguajes de programación, la unión de los lenguajes mencionados permite el desarrollo eficaz de páginas Web dinámicas con posibilidad de conexión a gestores de bases de datos y generación de gráficos personalizados, entre otras capacidades.

HTML (HyperTextMarkupLanguage) es el lenguaje de programación utilizado para crear las páginas Web. Básicamente se trata de una especie de editor de texto, donde se utilizan etiquetas o Tags, para formatear el texto, imágenes, entre otras. HTML tiene algunas funcionalidades diferentes a las que presenta un editor de texto, como los enlaces o Links que permite saltar de una página a otra con un simple click.

PHP (HyperText Preprocessor) es un lenguaje de programación incrustado en documentos HTML, que permite la navegación dinámica de contenidos en un servidor Web. Entre las características principales de PHP se destacan: La ejecución en el servidor, la disponibilidad de librerías de conexión con la mayoría de los sistemas gestores de bases de datos, la disponibilidad librerías

que manejan el entorno gráfico como jpgraph.php que permiten generar diferentes tipos de gráficos.

JAVASCRIPT es el lenguaje de secuencia de comandos en cliente más utilizado actualmente en la Web (Powell y Schneider, 2002). JAVASCRIPT proporciona una gran variedad de capacidades dinámicas en el cliente, un problema importante con los modelos de objetos basados en el navegador es que cada empresa distribuidora decide qué características presentar al programador y cómo hacerlo. Para combatir las incompatibilidades, el Consorcio W3C propuso un estándar que crea una correspondencia entre un documento HTML o XML y la jerarquía de objetos del documento. Este modelo se llama Modelo de Objetos de Documento o DOM, por sus siglas en ingles *Document Object Model* (Powell y Schneider, 2002). En la aplicación Web desarrollada se utilizó el estándar DOM siempre que fue posible.

A continuación se describen las interfaces desarrolladas con las respectivas restricciones o procedimientos.

En la Figura 11, se presenta la interfaz para la validación de usuario. Esta interfaz debe garantizar que:

- Se ingresen todos los datos solicitados.
- La base de datos exista en el gestor de bases de datos de PostgreSQL.
- El usuario exista en el gestor de bases de datos de PostgreSQL.
- La contraseña ingresada corresponda al usuario ingresado.
- El usuario tenga acceso a la base de datos guardada en PostgreSQL.

Al cumplirse lo anterior, se permite el acceso a la interfaz para la selección de datos para el análisis de regresión, en caso contrario, permite intentarlo nuevamente.

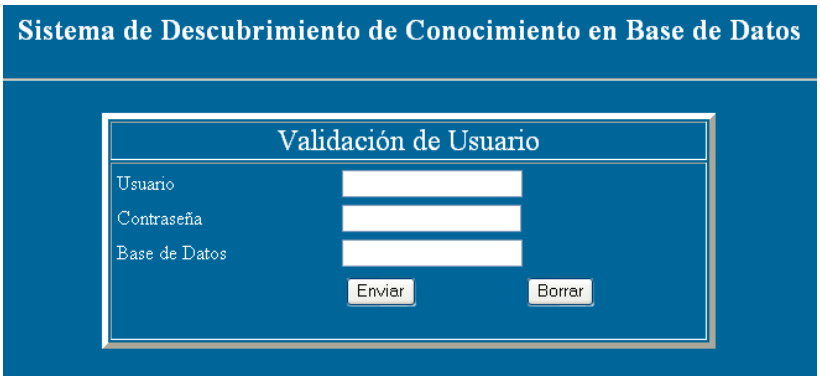


Figura 11. Interfaz para la validación de usuario

La Figura 12, presenta la interfaz de selección de atributos para el análisis de regresión, la interfaz esta diseñada para guiar al usuario en la selección de atributos para el análisis, por esto, presenta las siguientes características:

- Inicialmente solo se activa el primer campo “Origen/Tabla”, este campo permite seleccionar una tabla de la lista de tablas disponibles en la base de datos, excluyendo las tablas del sistema y las tablas estadísticas de las distribuciones F, T, χ^2.

- Después de seleccionar una tabla del campo “Origen/Tabla”, sólo se activa el campo “Atributo/Campo”.

- La lista de atributos que se presenta en el campo “Atributo/Campo” corresponde a los atributos de la tabla seleccionada en el campo “Origen/Tabla” de la misma fila.

- Solo se activan los demás campos cuando se selecciona un atributo en el campo “Atributo/Campo”.

- El sistema permite hacer una transformación al atributo seleccionado. Ejemplo: Utilizar para el análisis de regresión el logaritmo natural del atributo en vez del atributo. Se verifica que una transformación numérica requiere que el atributo sea numérico.

- Al final de la lista de transformaciones disponibles se encuentra “Variable Indicadora” que permite transformar cualquier tipo de dato en un valor de cero (0) o uno (1). Ejemplo: El atributo operador telefónico de tipo texto se puede utilizar como variable indicadora, con la condición “sí el operador telefónico es UNE utilizar el valor de uno (1), en caso contrario el valor de cero (0)”. En cualquier caso el atributo puede ser una variable indicadora que al cumplir una condición se le asigna el valor de uno (1), en caso contrario se le asigna el valor de cero (0).

- El cuarto campo es “Nombre Personalizado”, el sistema asigna el nombre del atributo seleccionado por defecto en este campo, pero el usuario puede asignarle cualquier nombre para identificar el atributo seleccionado en el resto del análisis.

- Los campos cinco y seis, permiten establecer una condición o filtro para definir o filtrar los datos con los cuales se realizará el análisis, en caso, de establecerse el atributo como una variable indicadora, la condición se utiliza para definir el valor a asignar.

- El campo “Tipo de Variable” permite al usuario establecer qué atributo es la variable respuesta, qué atributo es una variable explicativa o qué atributo no se incluye en el análisis (este último caso se presenta cuando el atributo sólo sirve para establecer un filtro). El sistema debe verificar que sólo se puede definir una variable respuesta.

- El sistema debe verificar inmediatamente la información ingresada por el usuario y guiarlo por medio de cuadros de diálogo con mensajes muy claros y específicos.

- En la interfaz se presenta un texto de ayuda que indica la acción a seguir en cada uno de los campos. Ejemplo: cuando el enfoque se encuentra en el campo “Atributo/Campo”, el texto de ayuda es “Atributo/Campo: Seleccione un campo de la lista de campos disponibles en la tabla seleccionada”, cuando el enfoque se encuentra en el último campo “Tipo de Variable”, el texto de ayuda es “Tipo de Variable: Indica si el atributo seleccionado es la variable respuesta o es una de las variables explicativas. Cuando no se incluye en el análisis es porque sólo sirve para establecer un filtro a los datos”.

- La creación de filas se realiza automáticamente cada vez que el usuario inicia la modificación de una nueva fila.

- La fila activa se resalta con un fondo de color azul claro para facilitar su identificación.

- No se permite editar o iniciar la edición de una nueva fila sin terminar de completar correctamente la fila activa.

Se establecen los atributos para el análisis en la tabla que se presenta en la Figura 12, se observa que el usuario puede seleccionar atributos de diferentes tablas de la base de datos, en estos casos, la interfaz presenta un botón de comando con el título “Establecer Relaciones entre Tablas” que permite hacer visible una tabla donde se definen las relaciones entre tablas. En la Figura 13, se ilustra un ejemplo donde se seleccionaron atributos de dos tablas, por lo tanto, se requiere definir la relación entre la tabla “autos” y la tabla “tabla_1”.

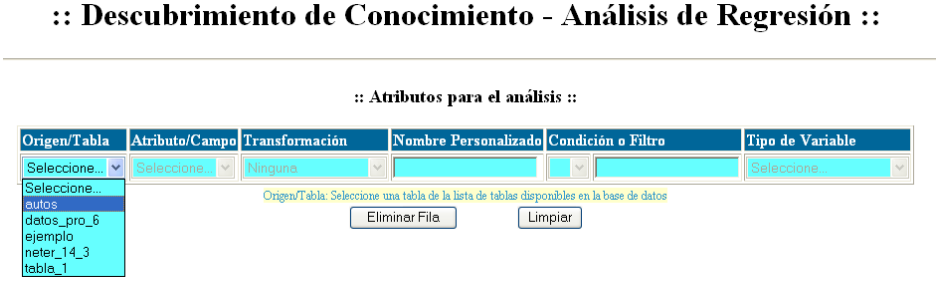


Figura 12. Interfaz para la selección de atributos

La interfaz permite establecer todos los tipos de relaciones disponibles en PostgreSQL, la tabla “Relaciones entre Tablas” también tiene un texto de ayuda al pie de la tabla, que presenta la descripción del tipo de relación seleccionado.

Figura 13. Ejemplo de selección de atributos de diferentes tablas

La Figura 14, presenta las demás opciones para el planteamiento del análisis de regresión. El “Tratamiento para los valores desconocidos” presenta dos opciones:

- Ignorar los valores nulos o valores desconocidos (Recomendado).
- Reemplazar los valores nulos por el valor promedio del atributo.

La interfaz permite al usuario establecer si desea incluir o no, el coeficiente independiente, también permite que el usuario seleccione el nivel de confianza con el cual se evalúan todas las hipótesis estadísticas y se determina la significancia de la regresión y de cada uno de los coeficientes de regresión. Al final, se establece el número de decimales con el cual se presentan los resultados.

Al dar clic en el botón “Iniciar Análisis”, se inicia una serie de procedimientos, de los cuales se enuncian los más importantes:

- Se verifica los atributos seleccionados, para el análisis debe existir, por lo menos, una variable respuesta y una variable explicativa.

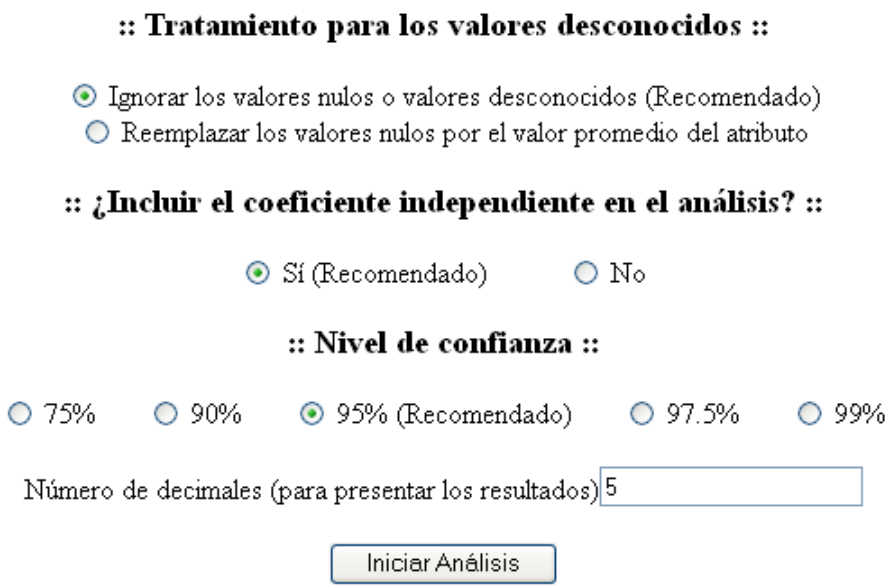


Figura 14. Opciones para el planteamiento del análisis

- El sistema define el tipo de análisis, dado que no todos los usuarios tiene por qué conocer las diferencias técnicas de los tipos de análisis, pero sí, le puede interesar las posibles relaciones existentes entre los atributos seleccionados como variables explicativas con la variable respuesta. Se observa que el usuario en ningún momento establece realizar un análisis de regresión lineal o un análisis de regresión logística, el tipo de análisis se define de una revisión de las características de la variable respuesta. Sí la variable respuesta es una variable tipo *Verdadero/Falso* o se establece como variable indicadora, la técnica a utilizar es la regresión logística, en caso contrario, se utiliza la técnica de regresión lineal.

- Con un cuidadoso procedimiento se construye la orden SQL correspondiente a la función a utilizar. El argumento más complejo para crear es el correspondiente al primer argumento que es una instrucción SQL para crear un vector con todos los datos para el análisis, para construir dicho argumento se realiza un análisis léxico, sintáctico y semántico. La construcción de este primer argumento se puede dividir por lo menos en cinco grandes pasos y cada uno de ellos requiere diferentes tipos de análisis para garantizar una regla bien formada. Se presenta a continuación una breve descripción y sin mayor detalle de los pasos en la construcción del primer argumento.

    1. La determinación de las diferentes tablas u orígenes en la selección de los atributos para el análisis.

2. La construcción de la regla o fórmula para especificar las relaciones de la cláusula FROM, considerando las relaciones establecidas y las tablas determinadas en el paso anterior.

3. La construcción de las condiciones para la cláusula WHERE, considerando los criterios o filtros establecidos y las relaciones que se expresan como condiciones.

4. La construcción de las sentencias necesarias para realizar las transformaciones planteadas por el usuario para cada uno de los atributos, la transformación a variable indicadora. La construcción de las sentencias necesarias para utilizar el valor promedio en vez de un valor nulo cuando el usuario así lo ha establecido. La instrucciones para convertir el tipo de dato en un tipo de dato *numeric* (tipo de dato proporcionado en PostgreSQL).

5. Generar una única fórmula o sentencia SQL bien formada, construyendo un vector con los valores de la variable respuesta concatenada con los vectores de los valores de las variables explicativas.

La Figura 15, presenta el planteamiento del análisis de regresión correspondiente a los resultados presentados en la Figura 6. La Figura 16, muestra la sentencia SQL generada por la interfaz inteligente para obtener el vector con los datos para el análisis de regresión. Se puede observar que con sólo unos pocos clics, sin escribir una sola palabra, se pueden seleccionar los datos para realizar cualquier análisis, en este caso, para analizar el *Rendimiento.* El proceso es intuitivo y no requiere que el usuario posea conocimientos en SQL. Si el análisis se planteara construyendo las sentencias SQL, pocos serían los usuarios que utilizarían el sistema para descubrir conocimiento.

La Figura 17, presenta la interpretación de los resultados del análisis planteado en la Figura 15. La presentación de resultados se enfoca a todo tipo de usuarios, en especial los gerentes, que desean la presentación del nuevo conocimiento en forma condensada, para esto, se utiliza el modelo de visualización descrito en el capítulo 4. Seguido del modelo de visualización, se presenta el modelo matemático que permite realizar predicciones, posteriormente, se presentan los comentarios sobre la significancia de la regresión y la significancia de las variables explicativas.

:: Atributos para el análisis ::

Origen/Tabla	Atributo/Campo	Transformación	Nombre Personalizado	Condición o Filtro	Tipo de Variable
autos	rendimiento	Ninguna	Rendimiento		Variable Respuesta
autos	potencia	Ninguna	Potencia		Variable Explicativa
autos	peso	Ninguna	Peso		Variable Explicativa
autos	aceleracion	Ninguna	Aceleracion		Variable Explicativa
Seleccione...	Seleccione...	Ninguna			Seleccione...

Descripción:

Eliminar Fila | Limpiar

Establecer Relaciones entre Tablas

:: Tratamiento para los valores desconocidos ::

(•) Ignorar los valores nulos o valores desconocidos (Recomendado)
( ) Reemplazar los valores nulos por el valor promedio del atributo

:: ¿Incluir el coeficiente independiente en el análisis? ::

(•) Sí (Recomendado) ( ) No

:: Nivel de confianza ::

( ) 75% ( ) 90% (•) 95% (Recomendado) ( ) 97.5% ( ) 99%

Número de decimales (para presentar los resultados): 5

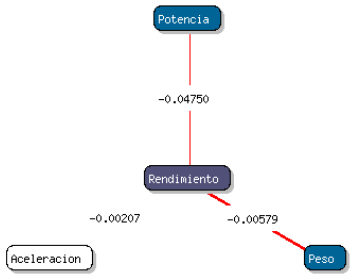
Iniciar Análisis

Figura 15. Planteamiento del análisis de regresión para describir la variable *Rendimiento*

```
array(select cast(autos.rendimiento as numeric) from autos where (autos.rendimiento is not null) and (autos.potencia is not null) and (autos.peso is not null) and (autos.aceleracion is not null)) || array(select cast(autos.potencia as numeric) from autos where (autos.rendimiento is not null) and (autos.potencia is not null) and (autos.peso is not null) and (autos.aceleracion is not null)) || array(select cast(autos.peso as numeric) from autos where (autos.rendimiento is not null) and (autos.potencia is not null) and (autos.peso is not null) and (autos.aceleracion is not null)) || array(select cast(autos.aceleracion as numeric) from autos where (autos.rendimiento is not null) and (autos.potencia is not null) and (autos.peso is not null) and (autos.aceleracion is not null))
```

Figura 16. Sentencia SQL para obtener el vector de datos

Representación Gráfica del Modelo



Coeficiente independiente = 45.67829

:: Modelo Matemático ::

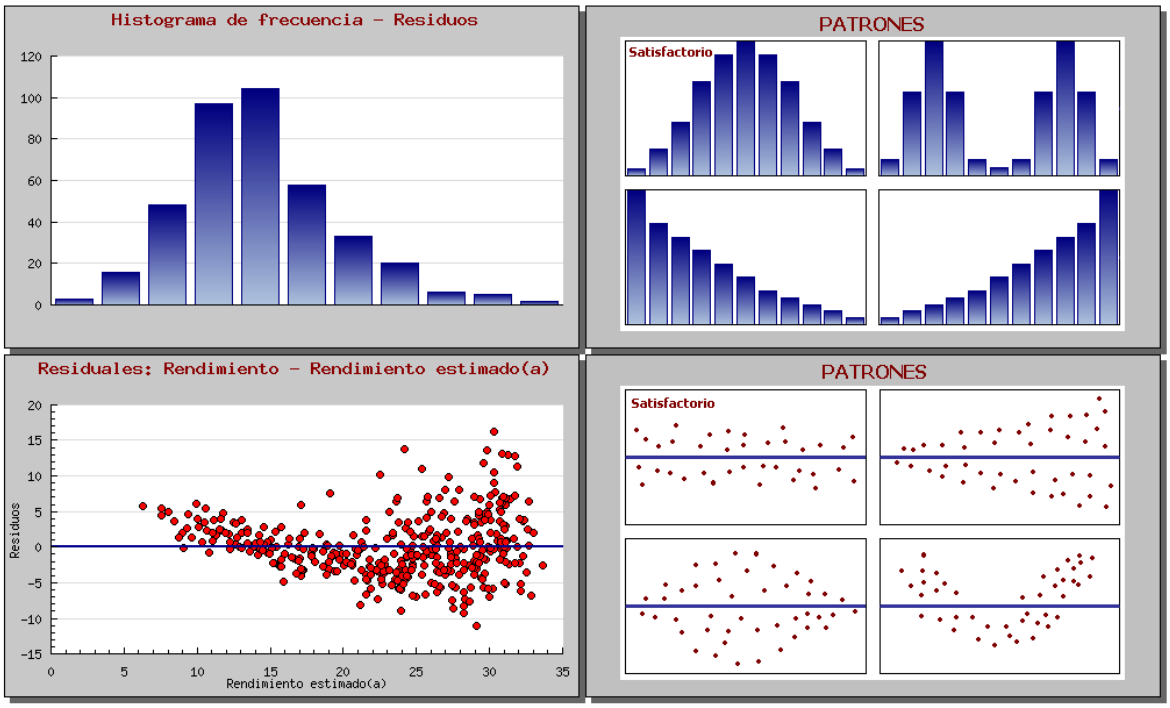
Rendimiento estimado(a) = (45.67829) + (-0.04750)*Potencia + (-0.00579)*Peso + (-0.00207)*Aceleracion

Se recomienda repetir el análisis de regresión sin considerar la variable: Aceleracion

Figura 17. Modelo de visualización y matemático para la variable *Rendimiento*

El prototipo construido puede presentar resultados adicionales correspondientes al mismo análisis, dado que todos los usuarios son diferentes y pueden estar interesados en diferentes resultados. Por lo anterior, se encuentran disponibles cuatro botones: "*Validación de supuestos*", "*Resumen del análisis*", "*Prueba de significancia*" y "*Matriz de varianzas y covarianzas*".

En el botón "*Validación de supuestos*" se presenta información gráfica y tabular que permite tener una idea del cumplimiento de los supuestos de la teoría de regresión lineal y logística multivariante, en la Figura 18, se presentan los resultados que se despliegan al dar clic en el botón.



Matriz de correlaciones simples entre variables explicativas

	Potencia	Peso	Aceleración
Potencia	1	0.86454	-0.68920
Peso	0.86454	1	-0.41684
Aceleración	-0.68920	-0.41684	1

ROJO: La multicolinealidad es significativa. AMARILLO: La multicolinealidad es moderada. VERDE: La multicolinealidad no es significativa.

Figura 18. Validación de supuestos para el modelo de *Rendimiento*

Los resultados presentados en la Figura 18, se pueden dividir en tres secciones, la primera es el histograma de frecuencia de los residuos con la gráfica de patrones para el histograma de frecuencia, la segunda es el gráfico de los residuales con la gráfica de patrones para el gráfico de residuales y por último la matriz de correlaciones simples entre variables explicativas.

El histograma de frecuencia sirve para comprobar que los errores tengan aproximadamente una distribución normal. La gráfica de residuales puede servir para verificar la condición de varianza constante o indicar si el modelo es inadecuado lo que puede sugerir alguna transformación para intentar estabilizar la varianza. La matriz de correlaciones simples entre variables explicativas es una primera aproximación para revelar los problemas de multicolinealidad.

En situaciones donde las dependencia entre las variables explicativas son fuertes, se dice que existe multicolinealidad (Montgomery y Runger, 2003). Existen varios métodos que pueden detectar la multicolinealidad, entre los más conocidos se encuentran: Los factores de inflación de la varianza, el determinante de la matriz de correlaciones simples y los valores propios de la matriz de correlaciones simples.

En el prototipo desarrollado se eligió presentar la matriz de correlaciones simples entre variables explicativas y permitir al usuario inspeccionar por sí solo cada elemento de la matriz. Los elementos de la diagonal siempre tienen el valor numérico de uno (1), para los demás elementos se debe verificar sí se aproximan a uno (1) en valor absoluto. En la literatura, no existe un consenso que permita determinar un valor concreto que se considere próximo a uno, y que indique que la multicolinealidad es un problema grave. Por lo anterior, en el presente trabajo se propone la siguiente escala que intenta revelar cuando la multicolinealidad es un problema grave que puede estar afectando la aplicabilidad general del modelo estimado.

Se define a r_{ij} como el elemento de la matriz que determina la correlación simple entre la variable explicativa i con la variable explicativa j.

- Sí $|r_{ij}| \geq 0.9$ se considera que la multicolinealidad es significativa y se resalta el elemento en color rojo.
- Sí $0.8 \leq |r_{ij}| < 0.9$ se considera que la multicolinealidad es moderada y se resalta el elemento en color amarillo.
- Sí $|r_{ij}| < 0.8$ se considera que la multicolinealidad NO es significativa y por lo tanto no es un problema, se resalta el elemento en color verde.

La inspección de cada elemento de la matriz de correlaciones simples entre variables explicativas no siempre permite detectar la presencia de la multicolinealidad, cuando más de dos variables de regresión están involucradas en una dependencia porque las r_{ij} no son necesariamente grandes (Montgomery y Runger, 2003).

En el botón "*Resumen del análisis*" se presenta información tabular con los coeficientes, los coeficientes estandarizados, el error estándar, el valor t, el

valor t de las tablas estadísticas y por último, se establece la significancia de la variable explicativa. En la Figura 19, se presentan los resultados que se despliegan al dar clic en el botón. La explicación del significado de los coeficientes de regresión se presentó en el capítulo 4.

Resultados del Ajuste del Modelo para: Rendimiento

Variables	Coeficientes	Coeficientes Estandarizados	Error Estándar	Valor t	T con alfa= (0.05/2)	Significancia
Coeficiente Independiente	45.67829					
Potencia	-0.04750	-0.23423	0.01599	-2.97049	1.974	Significativa
Peso	-0.00579	-0.63005	0.00058	-10.02379	1.974	Significativa
Aceleración	-0.00207	-0.00073	0.12334	-0.01675	1.974	NO Significativa

Figura 19. Resumen del análisis para el modelo de *Rendimiento*

En el botón "*Prueba de significancia*" se presenta la información relevante utilizada en la prueba de significancia del modelo, entre la cual se encuentran los grados de libertad, la suma de cuadrados, la media de cuadrados, el valor f, el valor f de las tablas estadísticas y por último, se establece la significancia de la regresión. En la Figura 20, se presentan los resultados que se despliegan al dar clic en el botón.

Prueba de Significancia de la Regresión

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Media de Cuadrados	Valor f	F con alfa= 0.05	Significancia
Regresión	3	16825.15309	5608.38436	311.13852	2.66	Significativa
Error o Residuo	388	6993.84038	18.02536			
Total	391	23818.99347				

Figura 20. Prueba de significancia para el modelo de *Rendimiento*

En el botón "*Matriz de varianzas y covarianzas*" se presenta información tabular con la matriz de varianzas y covarianzas. En la Figura 21, se presentan los resultados que se despliegan al dar clic en el botón.

Matriz de Varianzas y Covarianzas

	Coeficiente Independiente	Potencia	Peso	Aceleración
Coeficiente Independiente	5.80108	-0.02467	0.00040	-0.28045
Potencia	-0.02467	0.00026	-0.00001	0.00142
Peso	0.00040	-0.00001	3.33583E-7	-0.00004
Aceleración	-0.28045	0.00142	-0.00004	0.01521

Figura 21. Matriz de varianzas y covarianzas para el modelo de *Rendimiento*

Hasta el momento se han ilustrado algunos ejemplos de análisis con regresión lineal, pero el análisis de regresión logística se plantea de igual manera. Se debe recordar que el tipo de técnica a utilizar depende de las características de la variable respuesta, como se enunció anteriormente: Sí la variable respuesta

es una variable de tipo *Verdadero/Falso* o se establece como variable indicadora, la técnica a utilizar es la regresión logística. En la Figura 22, se presenta un ejemplo de selección de datos en el planteamiento para el análisis de regresión logística, en este caso, la variable respuesta es una variable indicadora binaria *Enfermedad*. Las variables explicativas son: *Edad, Socioeconómico 1 "SocEco1", Socioeconómico 2 "SocEco2"* y *Sector*. Los datos para el ejemplo están disponibles en el CD-ROM anexo al libro de Neter *et al.* (1996).

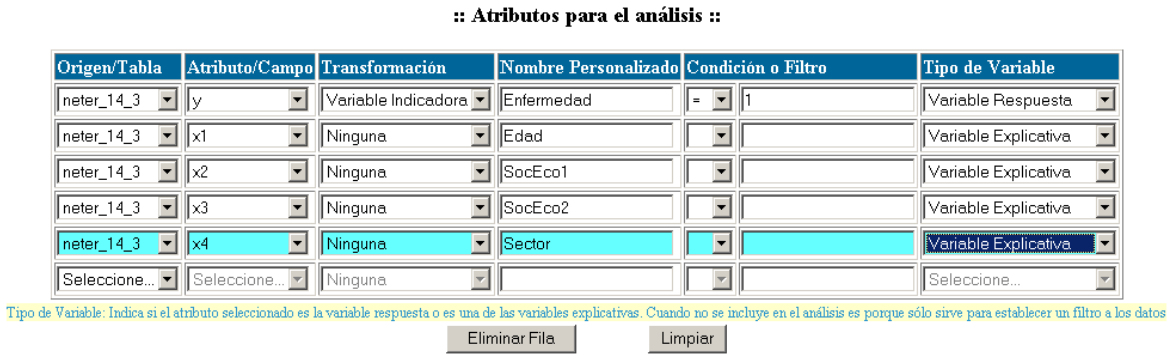
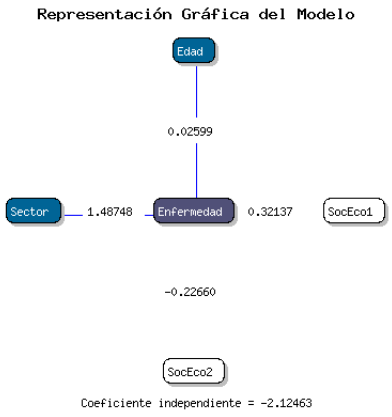


Figura 22. Planteamiento de un análisis de regresión logística para la variable *Enfermedad*

La Figura 23, presenta el modelo para la visualización de resultados y el modelo matemático del análisis de regresión logística multivariante planteado. Se observa que sólo las variables *Edad* y *Sector* son significativas.



:: Modelo Matemático ::

P(Enfermedad estimado(a) = 1) = [1 + exp(- (-2.12463) - (0.02599)*Edad - (0.32137)*SocEco1 - (-0.22660)*SocEco2 - (1.48748)*Sector]^(-1)

Se recomienda repetir el análisis de regresión sin considerar la variable: SocEco2

Figura 23. Modelo de visualización y matemático de la regresión logística para la variable *Enfermedad*

En la validación de supuestos para el análisis de regresión logística multivariante sólo se presenta información tabular que indique que la multicolinealidad es un problema grave, para ello se utiliza la matriz de correlaciones simples entre variables explicativas. La Figura 24, presenta los resultados de la validación de supuestos para el análisis de regresión logística multivariante para la variable *Enfermedad*. En el análisis de regresión logística no se presentan: el histograma de frecuencia de los residuos ni el gráfico de los residuales porque en general estas gráficas no siguen un patrón regular y pueden confundir al usuario.

Matriz de correlaciones simples entre variables explicativas

	Edad	SocEco1	SocEco2	Sector
Edad	1	-0.02550	-0.15816	0.15655
SocEco1	-0.02550	1	-0.43396	0.11872
SocEco2	-0.15816	-0.43396	1	-0.27358
Sector	0.15655	0.11872	-0.27358	1

ROJO: La multicolinealidad es significativa. AMARILLO: La multicolinealidad es moderada. VERDE: La multicolinealidad no es significativa.

Figura 24. Validación de supuestos de la regresión logística para la variable *Enfermedad*

La Figura 25, presenta el resumen del ajuste y prueba de significancia de la regresión logística multivariante para la variable *Enfermedad*.

Resultados del Ajuste del Modelo para: Enfermedad

Variables	Coeficientes	Error Estándar	"Deviance" Parcial	Error Cuadrático Parcial	Significancia
Coeficiente Independiente	-2.12463				
Edad	0.02599	0.01208	113.94484	19.30143	Significativa
SocEco1	0.32137	0.53146	101.87837	17.08340	NO Significativa
SocEco2	-0.22660	0.53989	101.86259	17.02364	NO Significativa
Sector	1.48748	0.45114	120.60191	20.63600	Significativa

Prueba de Significancia de la Regresión

Deviance	Número de Observaciones	Suma de Cuadrados del Error	Media de Cuadrados del Error	Ji-Cuadrada con alfa= 0.05	Significancia
101.26436	98	16.98114	0.18259	116.5	Significativa

Figura 25. Resumen del ajuste y prueba de significancia de la regresión logística para la variable *Enfermedad*

La Figura 26, presenta la matriz de varianzas y covarianzas de la regresión logística multivariante para la variable *Enfermedad*.

Matriz de Varianzas y Covarianzas

	Coeficiente Independiente	Edad	SocEco1	SocEco2	Sector
Coeficiente Independiente	0.35198	-0.00479	-0.14832	-0.16037	-0.14819
Edad	-0.00479	0.00015	0.00095	0.00055	0.00056
SocEco1	-0.14832	0.00095	0.28245	0.11689	0.01298
SocEco2	-0.16037	0.00055	0.11689	0.29148	0.05060
Sector	-0.14819	0.00056	0.01298	0.05060	0.20353

Figura 26. Matriz de varianzas y covarianzas de la regresión logística para la variable *Enfermedad*

Es común que se abuse del análisis de regresión, buscando determinar relaciones estadísticas entre variables que no están relacionadas desde el punto de vista práctico, en otras palabras, buscar determinar relaciones que no tienen sentido lógico. El prototipo desarrollado no puede evitar que el usuario plantee análisis de regresión sin sentido, pero si advierte al usuario cuando no es posible rechazar la hipótesis nula $\beta_1 = \beta_2 =,..., \beta_{p-1} = 0$, en la Figura 27, se presentan los resultados de un ejemplo en el cual no es posible rechazar la hipótesis nula.

:: Modelo Matemático ::

Y estimado(a) = (0.36842) + (0.04825)*X1 + (-0.17398)*X2

No existe suficiente evidencia estadística para asegurar que el modelo de regresión lineal es adecuado

Se recomienda repetir el análisis de regresión sin considerar la variable: X2

Figura 27. Resultados de un modelo de regresión inadecuado

7. CONCLUSIONES

La presente Tesis de Maestría se enmarca en un megaproyecto para la generación de un conjunto de herramientas gerenciales para facilitar y promover el Descubrimiento de Conocimiento en Bases de Datos, y por ende, facilitar y promover la aplicación de un enfoque de Inteligencia de Negocios.

El desafío de la popularización e implementación del enfoque de Inteligencia de Negocios en la toma de decisiones, consiste en facilitar el análisis, desarrollando e incorporando herramientas con la inteligencia suficiente para que colabore en la interpretación de la información derivada de la Minería de Datos. Con esto, se pretende disminuir la dependencia de especialistas a lo largo de todo el proceso de Minería de Datos, facilitando que un usuario final, sin dominio de alguna terminología específica o conocimientos técnicos, pueda por sí mismo realizar un análisis riguroso y con profundidad.

En el desarrollo de la investigación, se presentó un modelo para incorporar las técnicas de Minería de Datos, regresión lineal multivariante y regresión logística multivariante en un sistema gestor de bases de datos.

A continuación, se listan las principales conclusiones o aportes obtenidos en el desarrollo del trabajo investigativo:

- Se logró el acercamiento de las técnicas de Minería de Datos, regresión lineal y logística multivariante, para encontrar relaciones o dependencias entre variables, con usuarios no expertos en estadística o informática, y sin invertir dinero.

- Se logró la incorporación de las técnicas de regresión lineal y logística multivariante en un sistema gestor de bases de datos de distribución libre.

- Se planteó un modelo para la visualización de resultados de un análisis de regresión, que facilita la asimilación y utilización del nuevo conocimiento.

- Se construyó un prototipo para una aplicación Web que facilita la selección de los datos que deben ser incluidos en un análisis de regresión, por medio de una interfaz amigable e intuitiva.

- Se ha mostrado que el prototipo desarrollado interpreta por sí mismo, los resultados de un análisis de regresión, facilitando el Descubrimiento de Conocimiento en Bases de Datos a usuarios no expertos en estadística o informática.

- Se ha comprobado la factibilidad técnica de la incorporación de las técnicas multivariantes en un sistema gestor de bases de datos, por medio del diseño y construcción de un prototipo para una aplicación Web, que se utilizó para el análisis de regresión en varios ejemplos.

- El prototipo desarrollado presenta las características suficientes para brindar soporte a proyectos de Descubrimiento de Conocimientos en Bases de Datos.

Por los resultados obtenidos, el prototipo desarrollado se pretende aplicar en el proyecto de investigación "Descubrimiento de Conocimiento sobre la Innovación en Colombia a partir de las Encuestas de Innovación y Desarrollo Tecnológico, la Encuesta Anual Manufacturera y la base de datos ScienTI", financiado por COLCIENCIAS y ejecutado por varias universidades del país y el Observatorio Colombiano de Ciencia y Tecnología.

8. TRABAJO FUTURO

La presente Tesis de Maestría logró un acercamiento de las técnicas de Minería de Datos, regresión lineal y logística multivariante, para facilitar el descubrimiento de Conocimiento en Bases de Datos en usuarios no expertos en estadística o informática; lo que se considera un primer paso en el camino a la popularización del enfoque de Inteligencia del Negocio, quedando mucho por hacer. Por consiguiente, se listan posibles trabajos futuros:

- Ampliar las capacidades de un sistema gestor de bases de datos de distribución libre con la incorporación de otras técnicas de Minería de Datos diferentes a las técnicas de regresión lineal y logística multivariante.

- Permitir la inclusión de términos vagos para la selección de los datos y presentación de resultados, utilizando un lenguaje mucho más cercano al lenguaje natural.

- Construir una aplicación que permita plantear el análisis de regresión en forma similar o igual al modelo de visualización para la presentación de resultados propuesto en la presente Tesis de Maestría.

REFERENCIAS BIBLIOGRÁFICAS

Berger, C. y Haberstroh, B., (2005). “Oracle Data Mining 10g Release 2. Know More, Do More, Spend Less”. An Oracle White Paper. Oracle Corporation.

Buzan, T. y Buzan B. (1996). El Libro de los Mapas Mentales. España, Ed. Urano.

Chapman, P. y otros 6 autores (2000). “Metodología CRISP-DM para minería de datos. Guía paso a paso de Minería de Datos”. [En Línea], marzo de 2009, http://www.dataprix.com.

Chaudhuri, S. y Dayal, U., 1997. “An Overview of Data Warehousing and OLAP Technology”. Appears in ACM Sigmog Record, March 1997.

Draper, N.R. y Smith, H., (1966). Applied Regression Analysis. New York: Wiley.

Dumler, M., (2005). “Microsoft SQL Server 2005. Product Overview”. Microsoft Corporation.

Frawley, W., Piatesky-Shapiro, G. y Matheus, C. (1992), Knowledge Discovery in Databases. En: IA Magazine, Otoño, P. 213-228.

Hair, J.F., Anderson, R.E., Tatham, R.L. y Black, W.C., (1999). Análisis Multivariante. Madrid: Prentice Hall, 5 ed.

Han, J. y Kamber, M., (2001). Data Mining. Concepts and Techniques. Morgan Kaufmann Publisher. Academic Press.

Hand, D.J. (1989), Discrimination and Clasification. Ed. John Wiley & Sons.

ITL, (2006). Information Technology Laboratory, Statistical Reference Data Sets Archives. [En Línea], http://www.itl.nist.gov/div898/strd/general/dataarchive.html.

Jiménez, C. (2008), “Razonamiento Aproximado y Adaptable en el Procesamiento de Consultas Vagas”, Tesis Doctoral, Escuela de Ingeniería de Sistemas e Informática, Facultad de Minas, Universidad Nacional de Colombia, Sede Medellín.

Jiménez, C. (1999), “Modelos Conceptuales de la Ingeniería del software: Un análisis comparativo”, Tesis de Maestría, Facultad de Minas, Universidad Nacional de Colombia, Sede Medellín.

Larson, B., (2006). Delivering Business Intelligence with Microsoft SQL Server 2005. McGraw-Hill/Osborne, U.S.A.

Lockhart, T., (1999). Tutorial de PostgreSQL, Postgres Global Development Group. Editado por Thomas Lockhart, Libro electrónico.

Lopera, C.M., (2002). Identificación de grupos de datos influénciales en regresión logística mediante cluster. Medellín: Tesis de Maestría en Estadística. Universidad Nacional de Colombia.

Montgomery, D.C. y Runger, G.C., (2003). Applied Statistics and Probability for Engineers, 3rd Ed., John Wiley & Sons, Inc. ISBN 0-471-20454-4.

Mitra, S. y Acharya, T., (2003). Data Mining. Multimedia, Soft Computing and Bioinformatics. John Wiley & Sons. P. 5-7.

Neter, J., Kutner, M.H., Nachtsheim, C. y Wasserman, W., (1996). Applied Linear Statistical Models. McGraw-Hill, Fourth Edition. ISBN 0-256-11736-5.

Pérez, C., 2004. "Técnicas de Análisis Multivariante de Datos". Pearson Educación, S.A., Madrid.

Planeaux, D., Elumba, M. y Daniel, A., (2007). "New Features in Oracle Business Intelligence Suite Enterprise Edition 10g Release 3. An Oracle White Paper. Oracle Corporation.

Powell, T. y Schneider, F., (2002). JavaScript. Manual de referencia. McGraw-Hill/Interamericana de España, S. A. U., Madrid, Primera edición en español. ISBN: 84-481-3268-8.

Soto, C.M. y Jiménez, C., (2007). "Reto Informático para el Soporte de Decisiones en Bases de Datos Objeto-Relacionales". Presentado en el encuentro sobre Tendencias en Ingeniería de Software e Inteligencia Artificial, Medellín, 11-12 de Julio. ISBN: 978-958-44-1344-4.

Stackowiak, R., Rayman, J. y Greenwald, R. (2007). Oracle Data Warehousing and Business Intelligence Solutions. Wiley Publishing, Inc., Indianapolis, Indiana.

Utley, C., (2005). "Microsoft SQL Server 9.0 Technical Articles, Introduction to SQL Server 2005 Data Mining". [En línea], Libros en línea de Microsoft.

Weisberg, S., (2005). Applied Linear Regression. Third Edition. John Wiley & Sons, Inc., New Jersey.

Watt, A., (2006). Microsoft SQL Server 2005 For Dummies. Wiley Publishing, Inc. Indianapolis, Indiana.

Zvenger, P.A. y Fidel M., 2005. “Introducción al Soporte de Decisiones. Incorporación de Soluciones OLAP en entornos empresariales”, Tesis de Licenciatura, Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur.

ANEXOS (COPIA ELECTRÓNICA)

A. Código Fuente para PostgreSQL 8.2

La carpeta incluye el código necesario para incorporar las técnicas de regresión lineal y logística multivariante en cualquier computador o servidor que tenga instalado el gestor de bases de datos PostgreSQL 8.2 o una versión superior. Igualmente, se incluyen los datos de los ejemplos presentados. A continuación se enuncian los archivos contenidos en la carpeta:

- CREATE TABLE autos.sql
- CREATE TABLE pg_tb_dist_f.sql
- CREATE TABLE pg_tb_dist_t.sql
- CREATE TABLE pg_tb_dist_x2.sql
- FUNCTION pg_regresion_lineal.sql
- FUNCTION pg_regresion_logistica.sql
- FUNCTION pg_valor_f.sql
- FUNCTION pg_valor_t.sql
- FUNCTION pg_valor_x2.sql
- INSERT INTO autos.sql
- INSERT INTO pg_tb_dist_f.sql
- INSERT INTO pg_tb_dist_t.sql
- INSERT INTO pg_tb_dist_x2.sql
- SEQUENCE sec_autos.sql
- TYPE resultados_reg_lin.sql
- TYPE resultados_reg_log.sql

B. Código Fuente del Prototipo de la Aplicación Web

La carpeta incluye el código necesario para replicar la aplicación Web creada. A continuación se enuncian los archivos contenidos en la carpeta:

- index.html
- grafica_modelo.php
- histograma.php
- planteamiento.php
- residuos.php
- resultados.php
- validar.php
- Patron Histograma.bmp
- Patron Residuos.bmp

C. Tablas Estadísticas y Datos de los Ejemplos Utilizados

La carpeta incluye las tablas estadísticas de la distribución F, la distribución T y la distribución ji-cuadrada y los datos de ejemplo presentados en el documento. A continuación se enuncian los archivos contenidos en la carpeta:

- Datos Autos.xls
- Datos Ejemplo 14.3 - Libro Neter.xls
- Datos Problema 14.6 - Libro Neter.xls
- Tablas Estadísticas.xls

D. Archivos de Instalación de Postgres 8.2 (Distribución Libre)

Los archivos de instalación incluyen la herramienta PgAdmin III y toda la documentación de PostgreSQL.

E. Archivo de Instalación de WampServer 2 (Distribución Libre)

La herramienta WampServer 2.0 agrupa a Apache 2.2.8, PHP 5.2.5 e incluye la extensión “pgsql”.

F. Librerías para Graficar con PHP - JpGraph 2.3.3 (Distribución Libre)

La carpeta incluye las librerías de JpGraph-2.3.3 que permiten graficar con PHP y toda la documentación de la herramienta.

G. Artículo Publicado con Resultados Parciales del Trabajo Investigativo

El artículo “Reto Informático para el Soporte de Decisiones en Bases de Datos Objeto-Relacionales”, hace parte del desarrollo de la Tesis de Maestría y esboza el problema abordado en el trabajo investigativo.

Este artículo fue presentado en el encuentro realizado en la ciudad de Medellín sobre *Tendencias en Ingeniería de Software e Inteligencia Artificial*, 2007. ISBN: 978-958-44-1344-4.

H. Informe Ejecutivo

Los gestores de bases de datos son las principales herramientas para almacenar grandes cantidades de información. Estas herramientas ofrecen un lenguaje interactivo para plantear solicitudes de información. La mayoría de los gestores de bases de datos convencionales soportan el lenguaje estructurado de búsqueda SQL. Los sistemas de consulta-respuesta basados en SQL pueden responder a una gran cantidad de preguntas que representan gran parte del conocimiento almacenado en una base de datos; sin embargo, existe conocimiento oculto, representado en parte por relaciones que no son identificables a simple vista.

Tradicionalmente, la toma de decisiones se ha basado principalmente en juicios subjetivos, pero un nuevo enfoque gerencial toma cada vez más fuerza. A este enfoque gerencial se le denomina Inteligencia del Negocio, y se basa en la utilización de la información almacenada en bases de datos y de otras fuentes de información internas o externas, para tomar decisiones con diagnósticos más precisos y soluciones más inteligentes. En la actualidad, extraer conocimiento a partir de los datos para apoyar la toma de decisiones es indispensable en cualquier base de datos.

Las herramientas de Minería de Datos nacen de la necesidad del Descubrimiento de Conocimiento en Bases de Datos. La Minería de Datos se puede definir como un proceso analítico diseñado para explorar grandes cantidades de datos con el objetivo de detectar patrones de comportamiento o relaciones entre las diferentes variables. Los inconvenientes de utilizar las herramientas de Minería de Datos que ofrecen los gestores de bases de datos actuales y los paquetes estadísticos para el Descubrimiento de Conocimiento en Bases de Datos, se pueden resumir en:

- Se requiere de personal altamente calificado que domine la terminología de la Estadística o de la Inteligencia Artificial, manipulen las herramientas e interprete los resultados.

- Implica altas inversiones por el uso de herramientas comerciales, dado que las herramientas de distribución libre no tienen la robustez requerida para el manejo de grandes volúmenes de datos o la comunicación con los sistemas gestores de bases de datos.

- En el caso de los paquetes estadísticos, al ser herramientas independientes del almacenamiento de los datos, se requiere de tiempo para la preparación, importación o vinculación de los datos, prolongando así el tiempo de respuesta de los análisis y por ende su eficacia.

Para enfrentar el reto de eliminar o suavizar los inconvenientes arriba mencionados y facilitar el Descubrimiento de Conocimientos en Bases de

Datos, se incorpora las técnicas de regresión lineal y logística multivariante en un sistema gestor de bases de datos. El gestor de bases de datos de distribución libre utilizado fue PostgreSQL que se encuentra en la categoría de gestores de bases de batos objeto-relacionales. Adicionalmente, se presenta un modelo para la visualización de resultados que busca representar gráficamente los resultados esenciales del modelo estimado, basado en la interpretación de la regresión por medio de pruebas estadísticas de bondad de ajuste, tanto para la regresión como para los coeficientes de regresión.

Se desarrolló un prototipo de una aplicación Web para verificar la factibilidad técnica de la propuesta. La aplicación Web utiliza como servidor Web a Apache 2.2.8 y se desarrolla con la utilización de varios lenguajes de programación HTML, PHP 5.2.5 que incluye la extensión “PHP_pgsql” y JAVASCRIPT para el explorador Internet Explorer. La unión de los lenguajes mencionados permite el desarrollo eficaz de páginas Web dinámicas con posibilidad de conexión a gestores de bases de datos y generación de gráficos personalizados, entre otras capacidades.

Se demuestra que el prototipo desarrollado facilita la selección de los datos para un análisis de regresión. Adicionalmente, se implementa el modelo propuesto para la visualización de resultados y con ello, se obtiene una aplicación con la inteligencia suficiente para interpretar por sí mismo los resultados del análisis y presentarlos de manera amigable para apoyar la toma de decisiones, facilitando así, el Descubrimiento de Conocimiento en Bases de Datos a usuarios no expertos. Finalmente, el modelo conceptual para la incorporación de las técnicas de regresión multivariantes en un sistema gestor de bases de datos y el modelo para la visualización de los resultados, presentan características apropiadas para brindar soporte a proyectos de Descubrimiento de Conocimiento en Bases de Datos.